MODELLING MALARIA INCIDENCE IN THE LIMPOPO PROVINCE, SOUTH AFRICA: COMPARISON OF CLASSICAL AND BAYESIAN METHODS OF ESTIMATION

by

MAKWELANTLE ASNATH SEHLABANA

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. A Boateng (University of Cape Coast, Ghana)

CO-SUPERVISOR: Dr. D Maposa

2020

Declaration

I, **Makwelantle Asnath Sehlabana**, declare that the dissertation which is hereby submitted to the University of Limpopo, for the qualification of Master of Science in Statistics is my own independent work and has not been submitted by me for a qualification at this or any other institution of higher learning. Further, I have acknowledged all the sources used and listed these in the reference section.

Signature:.....Date:.....

Abstract

Malaria is a mosquito borne disease, a major cause of human morbidity and mortality in most of the developing countries in Africa. South Africa is one of the countries with high risk of malaria transmission, with many cases reported in Mpumalanga and Limpopo provinces. Bayesian and classical methods of estimation have been applied and compared on the effect of climatic factors (rainfall, temperature, normalised difference vegetation index, and elevation) on malaria incidence. Credible and confidence intervals from a negative binomial model estimated via Bayesian estimation-Markov chain Monte Carlo process and maximum likelihood, respectively, were utilised in the comparison process. Bayesian methods appeared to be better than the classical method in analysing malaria incidence in the Limpopo province of South Africa. The classical framework identified rainfall and temperature during the night to be the significant predictors of malaria incidence in Mopani, Vhembe and Waterberg districts of Limpopo province. However, the Bayesian method identified rainfall, normalised difference vegetation index, elevation, temperature during the day and temperature during the night to be the significant predictors of malaria incidence in Mopani, Sekhukhune, Vhembe and Waterberg districts of Limpopo province. Both methods also affirmed that Vhembe district is more susceptible to malaria incidence, followed by Mopani district. We recommend that the Department of Health and Malaria Control Programme of South Africa allocate more resources for malaria control, prevention and elimination to Vhembe and Mopani districts of Limpopo province. Future research may involve studies on the methods to select the best prior distributions.

Dedication

This dissertation is dedicated to my father and mother, Matome and Ngaletjane Sehlabana, who taught me the value of education and hard work. I also dedicate this work to my younger sister, Ditheto Sehlabana, who is looking up to me as her role model.

Acknowledgements

To the author and the finisher of my faith, the Almighty God, many thanks!
A very special gratitude goes out to my supervisor and co-supervisor, Dr A Boateng and Dr D Maposa, respectively. It was fantastic to have the opportunity to work on the entire research in your supervision, guidance, support and patience. Thank you very much. I am grateful to the following university staff: Prof M Lesaoana, Mrs A Ramalata and Mr D Mashishi for their unfailing support and assistance. I am also grateful to my siblings, father and mother, Ditheto, Tshegofatso, Kabelo, Morongwa, Matome and Ngaletjane, who provided me with moral and emotional support throughout this study. Thanks for all your encouragement! A special appreciation goes to the National Research Foundation (NRF) for funding my research.

Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

AIC	Akaike Information Criteria
ANOVA	Analysis of Variance
BLMs	Bayesian Linear Models
DCGE	Dynamic Compatible General Equilibrium
GLMs	Generalised Linear Models
GLMMs	Generalised Linear Mixed Models
HPD	Highest Posterior Density
MCMC	Markov Chain Monte Carlo
MCSE	Monte Carlo Standard Error
ML	Maximum likelihood
MLE	Maximum Likelihood Estimator
NB	Negative Binomial
NDVI	Normalised Difference Vegetation Index
StatsSA	Statistics South Africa
ZIP	Zero-Inflated Poisson
ZINB	Zero-Inflated Negative Binomial
ZTNB	Zero-Truncated Negative Binomial
ZTP	Zero-Truncated Poisson

Chapter 1

Introduction

1.1 Background of the study

Malaria is a mosquito borne disease caused by five protozoa, namely: Plasmodium Falciparum, Plasmodium Vivax, Plasmodium Malariae, related species of Plasmodium Ovale and Plasmodium Knowlesi (Snow, 2015). The protozoa are transmitted to humans through the bite of an infected female Anopheles mosquito (mosquito carrying protozoa) as illustrated in Figure 1.1. The Plasmodium Falciparum is known to have accounted for many malaria cases globally, and therefore, regarded as a threat to public health worldwide (Snow, 2015; Cox et al., 2018). Malaria incidence refers to the commonness of malaria. When the incidence rates are high, transmissions and prevalence of malaria are also high. This exposes the vulnerability and danger of the disease to the society. The symptoms of malaria include: fever $(> 37.5^{\circ}C)$, headache, rigors which are the repeated episodes of shivering, muscle pains, diarrhea, nausea, vomiting, loss of appetite, inability to feed babies, dizziness and sore throat. An example indicating how malaria is transmitted to humans is presented in

Fugere 1.1.

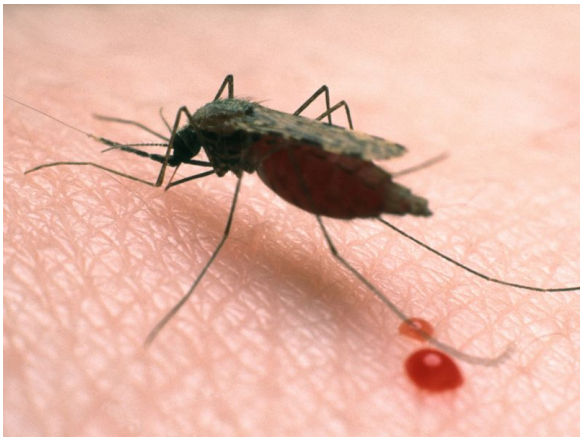

Figure 1.1: Malaria carrying mosquito

Based on history, malaria has infected and taken the lives of millions of individuals. This disease remains a major cause of human morbidity and mortality in most developing countries in Africa. Young children, pregnant women, and the elderly are the groups of people that still remain at high risk of malaria transmission (Schmidt, 2017). Sachs and Malaney (2002) outlined factors that contribute to increased malaria cases. These encompassed changing of agricultural practices, building of more dams, irrigation systems, deforestation, poor public health services and long-term climate change such as El Nino and global warming. Hay et al. (1998) found seasonal climatic change to be an important determinant of malaria incidence since variations in climate conditions could improve mosquito vector dynamics and parasite development rates (Najera et al., 1998). Indeed, malaria incidence has been found to be generally low during dry-hot season when vector populations are reduced and spatially restricted.

According to Blumberg and Frean (2017), there is a fairly good progress in malaria control globally. This progress is obtained through increased funding,

improved use of life-saving interventions and more countries pursuing malaria elimination. Although this progress was considerably achieved in countries such as Sri Lanka and some Sub-Saharan African countries, South Africa remains among the countries with high risk of malaria transmission (Raman et al., 2016), especially in the northern part of the country. Raman et al. (2016) further outlines that South Africa officially transitioned from controlling malaria to the goal of eliminating the disease in 2012. However, malaria cases have increased from 6811 in 2013 to 11,711 in 2014, with many cases reported in Mpumalanga and Limpopo provinces of South Africa.

1.2 Problem statement

Malaria remains a major cause of human morbidity and mortality in most developing countries. However, Africa is mostly affected by this disease. Young children, pregnant women, low immunity individuals, and the elderly are the groups of people at high risk of malaria transmission (Schmidt, 2017). Hay et al. (1998) found seasonal climatic change to be an important determinant of malaria incidence since variations in climate conditions could improve mosquito vector dynamics and parasite development rates (Najera et al., 1998). Malaria incidence is generally low during the dry-hot season when vector populations are reduced and spatially restricted. As a result, a number of researchers tend to focus on the peak transmission season. The season is often rainy. Hence the epidemiological picture during the dry-hot season is often neglected (Spottiswoode et al., 2014).

From the aforementioned reports, it is evident that malaria still remains a major health concern in the Limpopo province. There are consistent efforts meant to reduce malaria episodes. These include chemical spraying, use of treated mosquito bed nets, clearing bushes, cleaning drains and subsidised treatments,

and yet prevalence rates and malaria incidence remain high. It is probable that the efforts meant to reduce malaria menace do not specifically take into account the environmental factors likely to aggravate malaria disease. This study intends to fill this gap by incorporating these environmental factors into the generalised linear model estimated under classical and Bayesian approaches.

Again, most studies that modelled malaria cases, Boateng (2012), Omonijo et al. (2011), and Kleinschmidt et al. (2001) among others, employed classical methods such as Poisson, Negative binomial, hurdle, quasi-Poisson, dynamic computable general equilibrium (DCGE) models, to analyse the data. For these methods, the data may carry an uncertainty in the form of distribution of the sample. Moreover, the parameters from classical methods such as Poisson and negative regression models, have fixed population values such that the probability that an unknown parameter is any single value whose null hypothesis is always equal to zero, i.e, $\beta = 0$ (Zyphur and Oswald, 2015; Plonsky and Oswald, 2017). Due to the possible uncertainties and the assumptions preserved in the classical methods of estimation, there is a need to employ other methods that are different from the classical ones. This may help in assessing the results obtained through classical models to reveal their accuracy and reliability.

In this study, two statistical approaches are employed and compared: Classical and Bayesian estimation methods. The relation between the two statistical estimations results in the fact that the posterior distribution in the Bayesian approach is proportional to the likelihood function times the prior distribution. Whereas MLE uses asymptotic distributional assumptions in classical statistics, the uncertainty about model parameters in the Bayesian approach is expressed through the prior distributions. Combining the prior distribution and the likelihood (data), the researcher is able to update the knowledge about the model parameters. This is done via the posterior distribution from which

we can infer the estimates of the model parameters and relevant quantities like credible intervals.

1.3 Rationale

Malaria is one of the most severe public health problems worldwide. Therefore, it will be profitable to model malaria incidence in Limpopo province because it is among the provinces that account for most malaria cases in South Africa. This study is crucial because there are still arguments concerning the associations between environmental factors and malaria incidences. Yé et al. (2007) highlighted that the effects of climatic factors on malaria transmission are not efficiently assessed, specifically at local levels. Yé et al. (2007) also outlines that data used in many studies are proxy meteorological data obtained through satellites or interpolated from a different scale. This study will use local scale data from Malaria Control Institution in Limpopo province.

1.3.1 Aim of the study

The aim of the study is to determine the effect of environmental factors associated with malaria incidence by comparing classical and Bayesian estimation methods.

1.3.2 The objectives of the study

The objectives of the study are to:

i. Model malaria incidence given rainfall, temperature, normalised vegetation index, elevation and time in quarters from 2014 to 2015 across the various districts of the Limpopo province.

ii. Identify the effect of environmental factors which require more attention towards malaria control and prevention in Limpopo province.

iii. Examine the behavioral changes (trends) in overall malaria incidence in Limpopo province.

iv. Identify districts that are more susceptible to malaria incidence.

1.4 Methodology

The area of this study is composed of five districts of the Limpopo province: Mopani, Waterberg, Capricorn, Vhembe, and Sekhukhune. Malaria incidence data is provided by the Malaria Control Centre, based in Tzaneen, Limpopo province. Population data is provided by Statistics South Africa, and environmental factors (rainfall, temperature, elevation and normalised difference vegetation index (NDVI)) data were collected from Eco Verb. The data were collected monthly from January 2014 to June 2015. We have organised the data through tables, graphs, and condensed it into few summary measures. This has been attained through the use of descriptive analysis methods to reveal the essential characteristics of the raw data of interest. This has been useful in the model section for data analysis.

A Poisson model has been developed using the counts of monthly malaria incidences and each of the covariates (elevation, temperature during the night, temperature during the day, normalised difference vegetation index and rainfall). Over dispersion has been discovered in the developed Poisson model, therefore, the study has further employed a negative binomial model. Other models that can be employed include extra variation Poisson model, zero-inflated model or hurdle model. These models account for over dispersion naturally.

An additional model has also been developed and estimated via Bayesian estimation using the same data as the GLMs. A Bayesian estimation is based on Bayes theorem. The theorem helps in finding the shape of the posterior

distribution of the model. This distribution is crucial in Bayesian estimation. The results of this method have been compared with the results of the classical models which employed a maximum likelihood estimation, in order to reveal a better method of estimation. R software packages have been utilised for data analysis.

1.5 Ethical considerations

This study makes use of secondary data. Therefore, does not involve interaction with human samples. Hence there are no ethical issues which have been taken care of before utilising the data.

1.6 Significance of the study

The Limpopo Department of Health malaria control program will use the results of this study to assess the risk of malaria at the district level. The outcomes of the study will provide the South African malaria control programs, health policymakers and the Department of Health with an influential overview of malaria situation in Limpopo province. This overview will also help the government in planning and strategising for productive interventions for districts with a high risk of malaria. The results of this study could be used together with the results of related studies by other researchers to undertake further studies complementary to this one.

The Department of Health and other malaria control programs will also learn from this data analysis in order to adjust resource allocations regarding malaria prevention, control and elimination. Furthermore, the best method of estimation between classical and Bayesian approaches have been determined. This could help other researchers willing to undertake further similar studies.

1.7 Scope and limitations of the study

The major advantage of the study is the availability of resources, more specifically secondary data on the subject matter. As this study is novice and much previous research work have been conducted on this current topic, secondary data is available. Furthermore, government publications and online reference material have been relied upon for reviewing former studies that are related to the present study. The dissertation is only limited to its objectives. That is, the methodology of the dissertation is applied to model malaria incidence in the Limpopo province only.

1.8 Study area profile

Limpopo is a province in the northern part of South Africa. The province shares borders with three countries: Botswana, Mozambique, and Zimbabwe. The province is named after one of the most important rivers in the region, the Limpopo river. It is located on the border of Botswana and Zimbabwe. Limpopo province consists mainly of rural communities, with various ethnic groups and cultures. It is divided into five districts: Capricorn, Mopani, Sekhukhune, Vhembe and Waterberg. Limpopo province is located close to Johannesburg as the Braamfontein Spruit and the Crocodile River, before joining the Pienaar's River just after the Hartbeespoort Dam. Limpopo is surrounded by many rivers, dams and other kinds of water bodies. The map of the Limpopo river basin is presented in Figure 1.2.

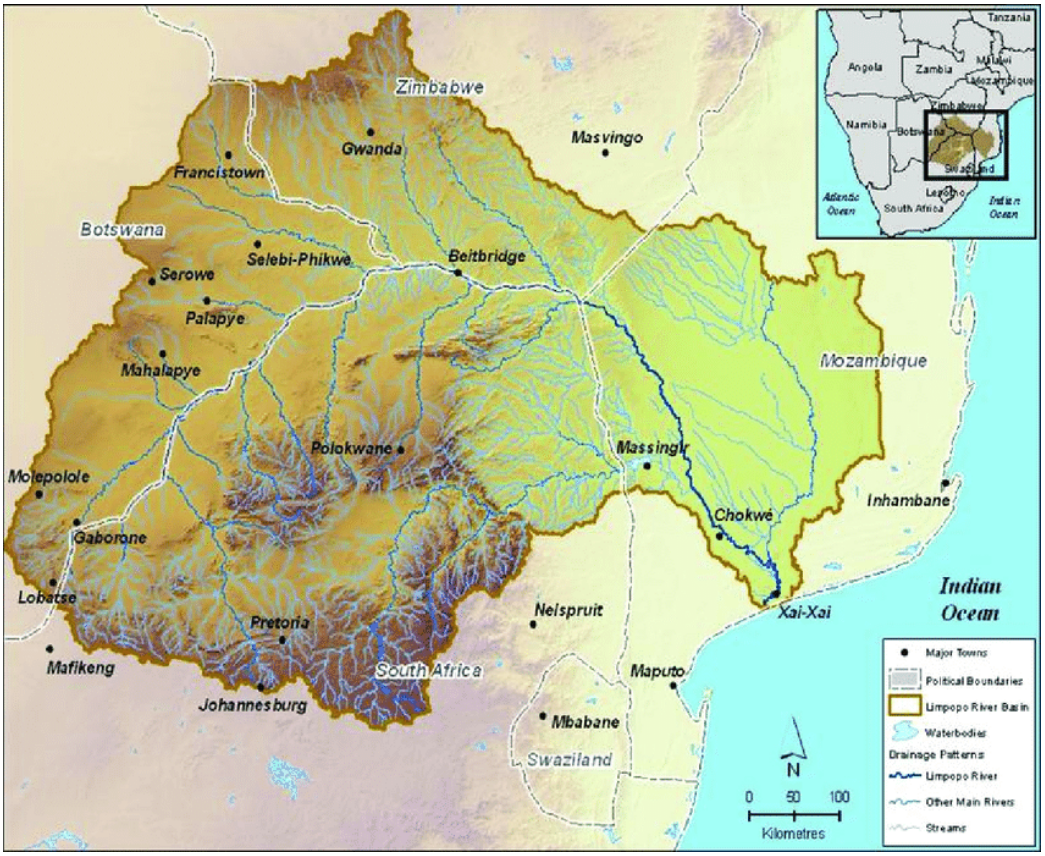

Figure 1.2: The Limpopo river basin

1.9 Organisation of the dissertation

The dissertation consists of five chapters. Background of the study, the study area profile, aim and objectives, problem statement, rationale, methodology, ethical considerations, significance of the study, scope and limitations of the study are discussed in Chapter 1. Chapter 2 scrutinises the related previous researches. This includes the exploration of various methodologies utilised in various researches to model malaria incidence. The methodology employed in the dissertation is discussed in Chapter 3. This includes the count data models in both classical and Bayesian frameworks. Chapter 4 comprises of data analysis and results explorations. Conclusions, recommendations and further

studies are examined in Chapter 5.

Chapter 2

Literature review

2.1 Introduction

This chapter surveys articles, dissertations, and other resources which are relevant to the study of modelling malaria incidence. The studies reviewed are conducted worldwide, including South Africa. Results from various studies are explored and summarised at the end of the chapter and relationships among the researcher's work are determined.

2.2 Related studies worldwide

Kazembe (2007) conducted a study titled "Spatial Modelling and Risk Factors of malaria incidence in northern Malawi". The study used regression models to find the spatial deviation of malaria risk. These regression models were also used in analysing the relationships between the risk of malaria and en-

vironmental factors at sub-district level in north Malawi. The environmental factors included altitude, precipitation and water holding capacity. Their study used monthly malaria case data collected between January 2002 and December 2003. Bayesian and Poisson regression models were employed with the assumption that spatial structures are different. The results of the study revealed that the covariates (environmental factors) were all significant. In addition, the results highlighted a positive relationship between malaria risk and the two covariates, altitude and precipitation. Areas of increased malaria cases were also identified for further epidemiological investigations.

The study by Shimaponda-Mataa et al. (2017) modelled the influence of temperature and rainfall on malaria incidence in Zambia, focusing on four malaria endemic provinces. Monthly data on malaria morbidity were analysed. Their data was collected for the period 2009 to 2012. The effects of these two covariates were modelled through a semiparametric Poisson regression model. The results of their study exhibited a strong positive association between malaria incidence and precipitation as well as minimum temperature.

Zayeri et al. (2011) coordinated a study to provide the geographical map of malaria and to identify some of the important environmental factors associated with malaria disease in Sistan and Baluchistan provinces of Iran. The data used to attain the results in accordance with the aim and objectives of their study, were a registered nine-year time series recorded from 2001 to 2009. The analysis part of the study was divided into two parts. The first part was a geographical mapping of malaria incidence rate and the second part was modelling the environmental factors. The former part employed empirical Bayesian estimation method and the latter part employed Poisson random effect for modelling of malaria incidences. The results of their study revealed that a large number of new malaria cases were observed between 2001 and

2009. Among those new cases, males accounted for many malaria cases compared to females. It was also discovered, based on their results, that adults (people older than 15 years) were more susceptible to malaria transmission as compared to children (aged younger than 5 years). Their results also indicated that rural areas had high malaria incidences than urban areas. Furthermore, they provided evidence that increase in humidity, elevation and temperature resulted in an increase in malaria incidences. However, increase in annual rainfall resulted in a decrease in malaria incidences. This negative association between rainfall and malaria cases is reported to be in contrast with the results of other related studies(e.g (Gosoniu et al., 2006) and (Kazembe, 2007)).

According to the study by Yé et al. (2007), climatic factors are associated with malaria transmission. Their study outlined that the effects of these factors are not efficiently assessed, especially at local levels. However, most of the studies aimed at assessing this association utilised proxy meteorological data obtained through satellites or interpolated from a different scale. The study by Yé et al. (2007) also addressed the relationship between the meteorological factors measured at the local scale and malaria infection. They selected a random sample of 676 children at the same time and scale between 1 January 2003 and 30 November 2004. The sample included children between 6-59 months. Data on some of the factors that can affect the incidence of malaria were also analysed. The covariates included temperature, humidity and rainfall in each site. These variables were measured monthly by digital meteorological stations. Their study employed logistic regression to predict the risks of malaria based on historical data. The results of their study revealed that the covariates were all significant. However, temperature was discovered to be the best predictor of clinical malaria rates. The effect of humidity on malaria risk was also discovered to be influential than of temperature. The association between rainfall and malaria was found to be positive. The study suggested that systematic

monitoring of temperature and rainfall could produce early warning system of malaria transmission risks.

Gosoniu et al. (2006) conducted a study that modelled a geostatistical malaria risk data. The objectives of the study included identification of significant environmental predictors of malaria transmission to assess the relations of environmental diseases. The assessment is used for prediction of malaria risks. The study used data from surveys conducted in Mali between 1977 and 1995. Climatic and environmental data were also extracted from different sources. The study focused on children aged less than 10 years. Bayesian stationary and non-stationary models were used to analyse malaria survey data. This model fit and its predictions were attained as the basis of Markov Chain Monte Carlo simulation methods. The climatic and environmental factors treated as covariates are seasonal length, Normalised Difference Vegetation Index (NDVI), temperature, rainfall and water bodies. The results of the study revealed a positive relationship between NDVI, minimum temperature, distance from permanent water bodies and malaria risk. A negative relationship between malaria risk and maximum temperature was discovered. The relationship between rainfall and malaria risk was found to be linear.

The influence of weather and climate on malaria occurrence based on human-biometeorological methods in Ondo State, Nigeria was modelled in the study by (Omonijo et al., 2011). Meteorological and malaria dataset for the period 1998 to 2008 was used. The study utilised Poisson distribution and log link function to examine the relationship between each of the biometeorological parameters and clinical reported malaria cases. Poisson multiple regression models were developed to assess the association between the explanatory variables and malaria cases. The results of the study revealed a positive relationship between wind speed, air temperature and sea surface temperature and malaria

cases.

The occurrence and incidence of malaria cases including severe cases reported at Obuasi Government Hospital in Ghana, was modelled in a study by (Boateng, 2012). Poisson and Negative Binomial regression models were fitted and compared. The developed models revealed no relationship between malaria incidence and gender. However, severe malaria cases were found to be prevalent among children aged less than 5 years and older people aged 70 years or more. The study found the Negative Binomial regression model to provide a better fit to the data compared to the Poisson regression model.

2.3 Related studies in South Africa

Gerritsen et al. (2008) generated a study that aimed to provide an overview of mortality rate and malaria incidences in Limpopo province from 1998 to 1999 and also from 2006 to 2007. This overview was used to reveal the trend of malaria incidences and mortality rate over time. Malaria and mortality data used were collected from Statistics South Africa. These data included information about population gender, age and districts. Chi-square tests were used to identify the trends of malaria incidence, the mortality rate over time and case fatality by age group. According to their descriptive statistical analysis results, a downward trend of malaria incidence was observed over the years of study. The mean incidence rate was found to be higher in males than in females. The incidence rate was also found to be lower in children (0-4 years) and higher in adults (35-39 years). A wide variability between the incidence rate and districts was outlined. Vhembe district had the highest incident rate and Sekhukhune had the lowest incidence rate. The study recommended the need for better data over a range of epidemic prone settings.

The study by Ramalata (2017) analysed malaria risk factors in the Limpopo province of South Africa. The study employed Poisson and negative binomial regression models to fit the data. Through the use of goodness of fit tests, the study revealed that Negative Binomial regression model outperforms Poisson regression model. The study found explanatory variables: rainfall, temperature during the night, two districts of Limpopo (Mopani and Vhembe), and seasonal effects such as Quarter1 (January - March) and Quarter4 (October - December) to be significantly associated with malaria incidences.

Kleinschmidt et al. (2001) conducted a study that used Generalised Linear Mixed Models (GLMMs) in the spatial analysis of small-area malaria incidence rates in KwaZulu-Natal, South Africa. Their study examined the association between malaria incidence and climatic and environmental factors. This was attained through the employment of GLMM with a Poisson distribution, a logarithmic link function, and a corrected error structure. The results of the study indicated that higher winter rainfall and higher average maximum temperature are positively associated with malaria incidence. The results also identified a negative association between increasing distance from water bodies and malaria incidence.

2.4 Summary of the chapter

Shimaponda-Mataa et al. (2017); Kazembe (2007) and Gosoniu et al. (2006) identified a positive relationship between rainfall and malaria risk. However, the results of the study by Zayeri et al. (2011) contradicted the results of the studies by (Shimaponda-Mataa et al., 2017), (Kazembe, 2007) and (Gosoniu et al., 2006). The results of the studies by Shimaponda-Mataa et al. (2017) and Gosoniu et al. (2006) identified a positive relationship between minimum temperature and malaria risk. Gerritsen et al. (2008) and Zayeri et al. (2011)

provided evidence that adults are more susceptible to malaria transmission than children.

The studies by Shimaponda-Mataa et al. (2017); Omonijo et al. (2011); Zayeri et al. (2011) and Kleinschmidt et al. (2001) modelled the environmental factors and assessed their relationships with malaria incidence through the development of Poisson regression models. However, Ramalata (2017) highlighted that the Negative Binomial regression model fits malaria incidence data better than the Poisson regression model.

The present study is crucial because there are still arguments concerning the associations between environmental factors and malaria incidences. Yé et al. (2007) highlighted that the effects of climatic factors on malaria transmission are not efficiently assessed, specifically at local levels. Based on the studies reviewed, the effects of climatic change on malaria have controversies. This could be due to the fact that the data used in many studies are proxy meteorological data obtained through satellites or interpolated from a different scale. Our study will use local scale data from Malaria Control Institution in Limpopo province.

Chapter 3

Research Methodology

3.1 Introduction

This chapter describes the broad profound framework of the methods used in the study. Section 3.2 defines the area of study and outlines how the data used in the study were generated. Section 3.3 describes the classical models and section 3.4 describes Bayesian method of estimation.

3.2 Study area and data collection

3.2.1 Study area

South Africa is one of the most diverse and attractive countries in the world. It is located on the southern tip of the African continent. It is bordered by Botswana, Mozambique, Namibia and Zimbabwe. South Africa has enjoyable climate and temperature, with warm sunny days most of the year. The summers run from November to February. The country is characterised by hot

weather with afternoon thunderstorms. Winters are generally mild and dry. South Africa is divided into nine provinces: Eastern Cape, Free State, Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, Northern Cape, North West and Western Cape. The area of this study is Limpopo province, which consists of five districts: Capricorn, Mopani, Sekhukhune, Vhembe, and Waterberg.

3.2.2 Study frame and data collection

We have modelled malaria incidence in Limpopo province of South Africa. Malaria incidence data is provided by Malaria control center. This center is based at Tzaneen town. The population data were provided by StatsSA. Environmental factors (rainfall, temperature, elevation, and normalised difference vegetation index) data were collected from Ecoverb. The data were collected monthly from January 2014 to June 2015.

3.3 Classical models

3.3.1 Generalised linear models

Generalised Linear Models (GLMs) are generalisation of Classical Linear Models. GLMs extend the framework of classical linear models to variables that are not normally distributed. The special cases of GLMs include: linear regression, logit and probit models, analysis of variance models (ANOVA), multinomial response models for count data and some models used for survival data. Linear models of classical regression analysis exhibit scaling problems, which are the results of the linear regression model assumptions. The assumptions combined include the constancy of variance, additivity of systematic effects and the approximations that errors are normally distributed. The scaling problems exhibited by linear models of classical regression models are reduced by the introduction of GLMs. As a consequence of introducing the GLMs, the con-

stancy of variance and the assumption that errors are normally distributed become less effective in classical regression models. However, the relationship between variance and mean is important and must be known. GLMs treat the additivity of the systematic effects as the expected responses. A GLM is given by:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_{p-1} x_{i,p-1} + \epsilon_i, \tag{3.1}$$

and in similar form is given as:

$$\vec{Y} = \vec{X}^\top \vec{\beta} + \vec{\epsilon}, \tag{3.2}$$

where:
Y_i is the response variable for observation $i = 1, 2, ..., n$,
x_{ij} is the explanatory variables $j = 1, 2, ..., p-1$,
β_0 is the regression intercept.
β_k is the regression coefficients $k = 1, 2, ..., m$, and
ϵ_i is the standard error.
The class of GLMs is specified using the following three components:

1. The random component.

   The random component specifies the conditional distribution of the response variable $Y_i (i = 1, ..., n)$ given the values of the explanatory variables in the model. When the dependent response outcomes $(y_1, ..., y_n)$ follow a probability distribution belonging to the exponential family of probability distributions, GLMs become easy to work with.

2. The systematic component.

   The systematic component is a linear function of a linear predictor. It is

based on the predictor variables. The systematic component is denoted by:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_{p-1} x_{i,p-1} \tag{3.3}$$

$$= \vec{X}^\top \vec{\beta}.$$

The regressors are the pre-specified functions of the predictor variables. Therefore, they can take on different forms of data types such as qualitative predictor variables, transformations of quantitative explanatory variables, polynomials, dummy variables, interactions, etc.

3. The link function.

   The link function, say g, relates the systematic component to the mean response. This is specified by:

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{i,p-1} \tag{3.4}$$

$$= \vec{X}^\top \vec{\beta}.$$

   The link functions in GLMs are said to be smooth and monotonic. Hence equation (3.1) can be written as:

$$\mu_i = g^{-1}(\eta_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{i,p-1} \tag{3.5}$$

$$= g^{-1}(\vec{x^{-1}}\vec{\beta}).$$

   If we choose $g = h$, where $\vec{\theta} = h(\vec{\mu})$, then:
   $\theta_i = h(\mu_i) = h(h^{-1}(\eta_i))$
   $= \eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{i,p-1}$
   $= \mathbf{X}^\top \vec{\beta}.$
   The link function in equation (3.2) is referred to as the canonical link.

3.3.2 Poisson Regression Model for count data

Count responses are different from other discrete responses. This is due to the fact that count responses cannot be expressed in the form of several proportions. The range of count data is theoretically unbounded and the upper limit of its number is infinite. The Poisson distribution plays an important role in modelling count responses. The importance of the Poisson distribution to count data is similar to the importance of the Normal distribution to continuous variables. Suppose that a random variable Y follows the Poisson distribution with parameter λ, then the Poisson distribution will be given by:

$$f\left(Y\right)\begin{cases}\frac{\lambda^{y}\exp^{-\lambda}}{y!}, \lambda > 0, y = 0, 1, 2, ..., \\ 0, \text{otherwise}.\end{cases} \tag{3.6}$$

Equation (3.6) is determined by one parameter λ, which represents both the mean and the variance of the distribution. There is no guarantee that all count variables will always follow a Poisson distribution. Hence it is always important to test whether or not a count variable satisfies the conditions of a Poisson law. This test is attained by combining the response that is larger than some threshold into a single category. This results into a multinomial variable. Therefore, the procedures for testing multinomial distributions are employed. These tests are used to determine whether the Poisson model is appropriate for describing the distribution of the original count variable. The procedure is carried out as follows: Suppose $\{y_i \colon 1 \leq i \leq n\}$ are the count observations from a sample of size n. Let m be the cut-point for grouping all responses $y_i \geqslant m$ and describe the count in cell n_j for the multinomial model obtained as follows:

$$n_j = \begin{cases}\text{number of } \{i \colon y_i = j\} \text{ if } 0 \leq j \leq m-1, \\ \text{number of } \{i \colon y_i \geqslant j\} \text{ if } j = m.\end{cases} \tag{3.7}$$

We determine the probability for each value of the response y_i under the null hypothesis of a Poisson distribution model as follows:

$$p_j = \begin{cases} f\left(j \mid \lambda\right), & 0 \leq j \leq m-1, \\ \sum_{y \geqslant m} f(j \mid \lambda), & j = m, \end{cases} \tag{3.8}$$

where $f\left(. \mid \lambda\right)$ is the Poisson distribution under the null hypothesis. The parameter λ changes from one observation to another. Therefore, this Poisson distribution can no longer be used to address the variation in λ. The Poisson regression model, commonly known as the Poisson log-linear regression model is the extension of the Poisson distribution to account for such heterogeneity. In Poisson log-linear model, the logarithm of the parameter λ is modelled, hence the name Poisson log-linear model. The logarithm of λ is treated as a linear function of explanatory variables. The Poisson regression model is a special case of GLMs. We have count responses denoted by $\vec{Y} = (y_i, ..., y_n)$, the explanotory variables denoted by $\vec{X_i} = (x_{i1}, x_{i2}, ..., x_{ip})^\top$ from the i^{th} subject $(1 \leq j \leq n)$. The Poisson regression model is then specified as follows:

1. The random component.

   Given the $\vec{X_i}$, the response variable $y_i \sim Poisson\left(\mu_i\right), i = 1, 2, ..., n.$

2. Systematic component.

   The conditional expectation of the response y_i given the explanatory variables $\vec{X_i}$ is linked to the linear predictor by the function of the logarithm of μ_i as follows:

$$\log\left(\mu_i\right) = \vec{X_i}^\top \vec{\beta} = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}, \tag{3.9}$$

   where:

   $\vec{\beta} = (\beta_1, ..., \beta_p)^\top$ is the vector of parameters of interest.

3. The link function.

   If $x_{i1} \equiv 1$, then β_1 will be the intercept. This Poisson log-linear model can model the variation in the mean of a count response that is explained by a vector of covariates. The Poisson distribution is a member of the exponential family and the log function in (3.4) is the canonical link for the Poisson model.

Parameter Interpretation

Let us first consider the case where x_i in Poisson regression model in (3.9) is an explanatory variable and β_1 is the coefficient of the covariate. The mean response for $x_1 = 1$ is given by:

$$E\left(y_i\right) = \exp\left(\beta_0 + \vec{\hat{x}}\vec{\hat{\beta}} + \beta_1\right), \tag{3.10}$$

where:
$\vec{\hat{X}}\vec{\hat{\beta}}$ is the vector $\vec{X}\vec{\beta}$ with the components of $x_1\beta_1$ removed. The mean response for $x_1 = 0$ is given by:

$$E\left(y_i\right) = \exp\left(\beta_0 + \vec{\hat{X}}^{\top}\vec{\hat{\beta}}\right). \tag{3.11}$$

The ratio of the mean responses for (3.10) and (3.11) is equal to $\exp\left(\beta_1\right)$. When x_1 is continuous, the mean response for $x_1 = a$ is given by:

$$E\left(y_i\right) = \exp\left(\beta_0 + \vec{\hat{x}}\vec{\hat{\beta}} + \beta_1 a\right). \tag{3.12}$$

Therefore, for each unit increase in the covariate x_1, the mean response is given by:

$$E\left(y_i\right) = \exp\left(\beta_0 + \vec{\hat{x}}\vec{\hat{\beta}} + \beta_1\left(a+1\right)\right). \tag{3.13}$$

Equation (3.13) is valid provided that the remaining components in the model are fixed. This implies that the mean response per unit increase in x_1 will be given by:

$$E(y_i) = \frac{\exp\left(\beta_0 + \vec{\hat{x}}\vec{\hat{\beta}} + \beta_1(a+1)\right)}{\exp\left(\beta_0 + \vec{\hat{x}}\vec{\hat{\beta}} + \beta_1 a\right)} = \exp(\beta_1).$$

When β_1 is positive, higher values of x_1 return higher mean responses, given that all the other covariates are fixed. When β_1 is negative, higher values of x_1 return lower mean responses, provided that other covariates are fixed. When $\beta_1 = 0$, the response y_i is independent of x_1. Therefore, to test whether x_1 is significant is equivalent to testing whether its coefficient is 0. The coefficient β_1 will generally change under a different scale of x_1. However, inferences such as those based on p-values about whether a coefficient is 0 remains the same, regardless of the scale used.

Inference about the model parameters

In this study, the method of maximum likelihood is used to estimate $\vec{\beta}$ for the Poisson log-linear model used in this study. The log-likelihood function is denoted by:

$$\ell\left(\vec{\beta}\right) = \sum_{i=1}^{n} \{y_i\mu_i - \exp(\mu_i) - \log(y_i!)\} \tag{3.14}$$

$$= \sum_{i=1}^{n} \left\{ y_i \vec{X_i}^\top \vec{\beta} - \exp\left(\vec{X_i}^\top \vec{\beta} - \log(y_i!)\right)\right\}.$$

Hence, the score function is given by:

$$\frac{\partial}{\partial \vec{\beta}} \ell\left(\vec{\beta}\right) = \sum_{i=1}^{n} \left\{ y_i \vec{X_i}^\top - \exp\left(\vec{X_i}^\top \vec{\beta}\right) \vec{X_i}^\top \right\}. \tag{3.15}$$

Now, noting that the second order derivative is negative, we have:

$$\frac{\partial^2}{\partial \vec{\beta}^\top} \ell\left(\vec{\beta}\right) = -\sum_{i=1}^{n} \exp\left(\vec{x}^\top \vec{\beta}\right) \vec{X}_i \vec{X}_i{}^\top < 0. \tag{3.16}$$

Equation (3.15) is trivial. Therefore, the MLE of $\vec{\beta}$ is well defined. We also note from (3.16) that the MLE of $\vec{\beta}$ is asymptotically distributed normally with the mean $\vec{\beta}$ and variance $\frac{1}{n}\Sigma$, where $\Sigma = \mathbf{I}^{-1}\left(\vec{\beta}\right)$ and $I\left(\vec{\beta}\right)$ is the Fisher information matrix. The asymptotic variance of the MLE of $\hat{\vec{\beta}}$ is given by $Var\left(\hat{\vec{\beta}}\right) = \frac{1}{n}\mathbf{I}^{-1}\left(\vec{\beta}\right)$, which is the inverse of expected Fisher information matrix. The Poisson model:

$$E\left(\mathbf{I}\left(\vec{\beta}\right)\right) = \left(\mu_i \vec{X}_i \vec{X}_i{}^\top\right),$$

where the Fisher information is specified by:

$$\mathbf{I}\left(\vec{\beta}\right) = \frac{1}{n}\sum_{i=1}^{n} \mu_i \vec{x}_i \vec{X}_i{}^\top.$$

Expressing $E\left[\mathbf{I}\left(\vec{\beta}\right)\right]$ in a closed form can be difficult since it depends on the distribution of x_i. Hence for inference purposes, the observed version of the Fisher information $\mathbf{I}\left(\vec{\beta}\right)$ is used with $\hat{\vec{\beta}}$ considered as an estimate of $\vec{\beta}$. We are interested in finding out if each covariate and the response variable are related. We can attain this by testing whether or not the coefficients of the covariates in the model are equal to zero. This method is more accurate when there is only one term involving the variable in the model and the variable is either continuous or binary. Another approach to this method is to use the MLE of $\vec{\beta}$ and its asymptotic normal distribution. If the variable is categorical and consists of many levels, dummy variables can be used to represent some of the variables in the model. Testing the relationship between the covariates and the response variable can also be attained through Wald, Score and Likelihood ratio tests.

Offsets in Poisson regression model

The observation period among the subjects of interest may vary in many studies. Hence these possible variations must be taken into account when the occurrence of count response is modelled. This accountability is important because the subjects with longer period of observation are more likely to have more events compared to subjects with shorter period of observation. Suppose we have a sample of size n and t_i is defined as the length of period of observation for the i^{th} subject. We also assume that the rate of event of the count response of interest follows the Poisson distribution. Therefore, the rate of this event can be modelled with Poisson regression model where the rate for the i^{th} subject is denoted by:

$$r_i = \exp\left(\vec{X_i}^{\top}\vec{\beta}\right).$$

When the period of observation (t_i) is different among the subjects, the count of event y_i for each individual subject i also follows a Poisson distribution with mean:

$$\mu_i = t_i r_i = \exp\left(\vec{X_i}^{\top}\vec{\beta}\right).$$

Therefore, we can still model the mean response μ_i using Poisson regression model as follows:

$$\log \mu_i = \log t_i + \log r_i \tag{3.17}$$

$$= \log t_i + \log\left[\exp\left(\vec{X_i}^{\top}\vec{\beta}\right)\right]$$

$$= \log t_i + \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_i p - 1.$$

When the period of observation is the same for all the subjects, that is, $t_i = t$, the $\log t_i$ is absorbed into β_0. This integrates (3.17) back to the Poisson regression model. When the period of observation t_i is different among the subjects, $\log t_i$ is taken as a covariate in the Poisson regression model. However, it is not

treated the same way as other covariates because its coefficient is always one. In GLM classification, $\log t_i$ is called the offset.

Goodness of fit

It is essential to check how good the model is in terms of fitting the data. This can be achieved by performing the goodness of fit tests. Among others, this study discusses only two goodness of fit tests that can be used to assess how good the Poisson regression model is in fitting the data.

1. Pearson's chi-square statistic

   Pearson's chi-square statistic is the sum of the normalised squared differences between the expected and the observed counts of the response variable. Under certain conditions, the Pearson chi-square statistic follows a chi-square distribution. As a result, the Pearson's chi-square statistic presents a goodness of fit test that is more reliable. Suppose that y_i is the count response and μ_i is the fitted value under Poisson regression model, where:

   $$\mu_i = \exp\left(\vec{X_i}^{\top}\vec{\hat{\beta}}\right).$$

   Then, this Poisson regression model is obtained by substituting $\vec{\beta}$ with $\vec{\hat{\beta}}$ in the mean response in (3.7). In Poisson distribution, the mean and the variance are equal. Therefore, we can estimate the variance by $\hat{\mu_i}$. So, the normalised squared difference for the i^{th} subject can be expressed as:

   $$\frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}.$$

   The Pearson's chi-square statistic is given by:

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

   where:

y_i is the number of subjects observed, which fall into the i^{th} pattern ($1 \leq i \leq n$), and $\hat{\mu_i}$ is the number of subjects expected, which fall into the i^{th} pattern $(1 \leq i \leq n)$. The Poisson distribution converges to a Normal distribution when the mean approaches infinity. Therefore, if the sampling variability of the estimate $\hat{\beta}$ is excluded, then:

$$\frac{\hat{y_i} - \hat{\mu_i}}{\sqrt{\hat{\mu_i}}} \sim N(0,1),$$

provided that $\hat{\mu_i} \longrightarrow \infty \ \forall \ 1 \leq i \leq n$. Hence for a fixed n, the Pearson statistic asymptotically follows a chi-square distribution with $n - p$ degrees of freedom. That is:

$$\sum_{i=1}^{n} \frac{(y_i - \hat{\mu_i})^2}{\hat{\mu_i}} \sim \chi^2_{n-p},$$

when $\hat{\mu_i} \longrightarrow \infty \ \forall \ 1 \leq i \leq n$ and p is the number of parameters to be estimated from the sample. Hence the following hypothesis can be tested:
H_0: the model is good.
H_1: the model is not good.

2. Scaled Deviance statistic

   The deviance statistic is defined as two times the difference between the maximum log-likelihood and the value of the log-likelihood obtained through the MLE method of the model parameter vector. Suppose that $\vec{Y} = (y_1, y_1, ..., y_n)\top$ is the response vector from the sample of size n. The deviance statistic of the model is defined as:

$$D(\vec{y}, \vec{\theta}) = 2\left[\ell(\vec{y}, \vec{y}) - \ell(\vec{y}, \vec{\theta})\right], \tag{3.18}$$

   where:
   $\ell(\vec{y}, \vec{y}) =$ the log-likelihood given that the model gave a perfect fit.

$\ell(\vec{y}, \vec{\theta}) =$ the log-likelihood for the model of interest.

Now, for the Poisson log-linear regression model, the deviance statistic is defined as:

$$D(\vec{y}, \vec{\theta}) = 2 \sum_{i=1}^{n} \left[y_i \log \left(\frac{y_i}{\hat{\mu}_i} \right) - (y_i - \hat{\mu}_i) \right], \tag{3.19}$$

where:

$\hat{\mu}_i = \exp\left(\vec{X_i}^{\top} \vec{\hat{\beta}} \right)$. When the Poisson log-linear regression model is correct, its deviance statistic in (3.19) follows a chi-square distribution with $n - p$ degrees of freedom. That is, $D(\vec{y}, \vec{\theta}) \sim \chi^2_{n-p}$. Similar to the method of using Pearson's chi-square, the following hypothesis can also be tested:

H_0: the model is good.

H_1: the model is not good.

The model is said to be good when there is no evidence of lack of fit in the model. Now, to assess how good the Poisson log-linear model is in terms of model fit, we divide the deviance statistic in (3.11) by the degrees of freedom $n - p$. If the resulting value is significantly larger than 1, then there is evidence of lack of fit. This approach of testing for goodness of fit can be used for both Pearson chi-square statistic and scaled deviance statistic.

Overdispersion in Poisson regression model

If the assumption of equality between the mean and the variance of the Poisson distribution is violated, the Poisson regression model becomes overdispersed. Observations that are based on time intervals of varying lengths and data clustering are responsible for overdispersion.

Detection of overdispersion

The two goodness of fit tests, Pearson's chi-square and scaled deviance statistics can be used to detect the chances of overdispersion occurrence. This is attained by dividing the statistic by the degrees of freedom $n - p$. When the resulting outcome is significantly greater than 1, then there are more chances of overdispersion in the model. Hence overdispersion is detected. According to some simulation studies, Pearson's chi-square is found to be the better method in detecting overdispersion (Hilbe, 2011).

Correction of overdispersion

There are two ways in which overdispersion can be corrected.

1. If the detected overdispersion in Poisson regression model is due to the observations that are based on time intervals of varying lengths, then the best method for correcting this kind of overdispersion is using the variance estimate to account for overdispersion. This method is not discussed in detail in this study.

2. If the source of overdispersion is data clustering, and the nature of this clustering is well understood, then the best method for correcting this kind of overdispersion is the development of the refined models specified in the next subsections.

3.3.3 Negative Binomial model

The Negative Binomial model is appropriate for the correction of overdispersion if the occurrence of overdispersion is due to the fact that observations are based on the unknown time intervals of different lengths. When developing this model, the mean (μ_i) of Poisson distribution is no longer treated as the parameter. However, the mean (μ_i) for each subject is treated as a random

variable. Heterogeneity is then allowed or introduced since a specific distribution has been defined for μ_i. For instance, let us assume that μ_i follows a Gamma distribution with a scale parameter θ_i and shape parameter α_i. This Gamma distribution is denoted by:

$$f(\mu_i) = \frac{(\mu_i)^{\alpha_i - 1} \exp(-\mu_i \theta_i / (1 - \theta_i))}{\Gamma(\alpha_i)((1 - \mu_i)/\mu_i)^{\alpha_i}}. \tag{3.20}$$

The Gamma distribution in (3.20) is integrated in order to obtain the marginal distribution for count data. This marginal distribution turns out to be the Negative Binomial distribution denoted by:

$$f\left(y_i \mid \theta_i \alpha_i\right) = \frac{\Gamma(y_i + 1/\alpha_i)}{y_i! \Gamma(1/\alpha_i)} \left(\frac{1}{1 + \alpha_i \theta_i}\right)^{1/\alpha_i} \left(\frac{\alpha_i \theta_i}{1 + \alpha_i \theta_i}\right)^{y_i}, \alpha_i > 0, y = 0, 1, ..., \tag{3.21}$$

where:

$\Gamma(\alpha_i)$ is the Gamma function,

α_i is the number of successes,

θ_i is the probability of success, and

y_i is the response variables.

The mean and variance of the Negative Binomial distribution are respectively given by

$$E\left(y_i \mid \mu_i \alpha_i\right) = \theta_i$$

and

$$var\left(y_i \mid \mu_i \alpha_i\right) = \theta_i (1 + \alpha_i \theta_i).$$

The Negative Binomial model is also a special case of the GLM. This model is specified by a systematic component given by:

$$\log\left(\mu_i\right) = \log\left[E\left(y_i \mid \vec{X_i}\right)\right] = \vec{X_i^\top} \vec{\beta}, 1 \leq i \leq n. \tag{3.22}$$

The variance of the Negative Binomial in (3.22) is always larger than the mean, provided that $\alpha_i \neq 0$. This adds a term $\alpha_i \theta_i^2$ to the variance of a Poisson distribution. This added quadratic term accounts for overdispersion in an overdispersed Poisson regression model, hence α_i is known as the dispersion parameter. As $\alpha_i \longrightarrow 0$ in (3.21), the NB distribution gets closer to the Poisson distribution. As α_i increases, the overdispersion is corrected.

Inference for the NB model:

We can make inference using the Maximum Likelihood (ML) because the NB log-linear model belongs to the GLMs family. The log-likelihood for this model is given by:

$$\ell\left(\vec{\beta}, \alpha_i\right) = \sum_{i=1}^{n} \log f_{NB}\left(y_i \mid \vec{X}_i, \vec{\beta}, \alpha_i\right) \tag{3.23}$$

$$= \sum_{i=1}^{n} \left\{ y_i! \left[\log g_1^{-1}\left(\vec{X_{1i}^{\top}}\vec{\beta}\right) - \log\left(\frac{1}{\alpha_i} + g_1^{-1}\left(\vec{X_{1i}^{\top}}\vec{\beta}\right)\right)\right]\right\}$$

$$+ \sum_{i=1}^{n} \left[\alpha_i \log\left(1 - \alpha_i g_1^{-1}\left(\vec{X_{1i}^{\top}}\vec{\beta}\right)\right) + \log\Gamma(y_i + 1/\alpha_i)\right]$$

$$- \sum_{i=1}^{n} \left(\log y_i! - \log\Gamma(1/\alpha_i)\right).$$

The MLE $\vec{\hat{\theta}}$ of $\vec{\theta} = (\vec{\beta^{\top}}, \alpha_i)^{\top}$ and its associated asymptotic distribution is obtained by maximising (3.23). To determine whether there is overdispersion in the data we can test for the stated hypothesis:
$H_0 : \alpha_i = 0$ vs $H_1 : \alpha_i \neq 0$ for $i \neq j$
However, we know that $\alpha_i \geqslant 0$ from (3.21). Hence, $\alpha_i = 0$ under H_0 is a boundary point. Therefore, the inference based on the asymptotic distribution of the MLE $\hat{\alpha}_i$ of α_i cannot be valid since 0 is not an inclusive point of the parameter space. In this case, inference about this boundary point can be based on a

modified asymptotic distribution. When using this approach to test for the null hypothesis ($H_0 : \alpha_i = 0$), the modified asymptotic distribution is a mixture of a point mass at 0 and the non-negative half of the asymptotic normal distribution of $\hat{\alpha}_i$ is gathered into a point mass centered at 0 naturally. This is due to the fact that negative values of α_i are not allowed under H_0.

3.3.4 Zero-inflated Poisson and Zero-inflated Negative Binomial models

In cases where overdispersion is caused by the presence of too many zeros in the data, the NB regression model does not correct overdispersion when it is employed. However, models that can be used to correct this kind of overdispersion are Zero-Inflated Poisson (ZIP) and Zero-Inflated NB (ZINB) models. These models account for structural zeros, better known as excess zeros, which are found within the data of interest. ZIP and ZINB models use the mixture distribution notations. Since ZIP and ZINB are closely related, in this study we will only discuss the ZIP regression model. This model is based on a mixture of Poisson distribution with parameter μ and a degenerate distribution of a constant 0. The mixture distribution is given by:

$$f_{ZIP}\left(Y \mid \rho, \mu\right) = \rho f_0(y) + (1-\rho) f_p\left(y \mid \mu\right), y = 0, 1, ... \tag{3.24}$$

where:
$f_0\left(y\right)$ is the probability distribution function of a constant 0, and $f_p\left(y \mid \mu\right)$ is the Poisson distribution with the parameter μ. Consider the distribution of a constant 0, that is, $f_0(0) = 1$ and $f_0(y) = 0, \ \forall \ y \neq 0$. Hence we can write (3.24) as:

$$f_{ZIP}\left(Y \mid \rho, \mu\right) = \rho f_0\left(y\right) + (1-\rho)\, f_p\left(y \mid \mu\right), y = 0, 1, ... \tag{3.25}$$

where:

$f_0(y)$ is the probability distribution function of a constant 0, and $f_o(y \mid \mu)$ is the Poisson distribution function of μ. Consider the distribution of a constant 0 with mass point at 0, and $f_0(y) = 0, \ \forall \ y \neq 0$. Hence we can write (3.25) as:

$$f_{ZIP}(y \mid \rho, \mu) = \begin{cases} \rho + (1-\rho) f_p(0) & if \ \ y = 0, \\ (1-\rho) f_p(y \mid \mu) & if \ \ y > 0. \end{cases} \tag{3.26}$$

Therefore, at $y \leq 0$, the Poisson distribution $f_p(0 \mid \mu)$ is inflated by ρ to address the excess zeros in the model. In this case the mixture distribution is given by:

$$f_{ZIP}(0 \mid \rho, \mu) = \rho + (1-\rho) f_p(0 \mid \mu). \tag{3.27}$$

The mixture distribution $f_{ZIP}(0 \mid \rho, \mu)$ must be constrained between 0 and 1 because of the probability at $y \leq 0$, which is the probability for zero. Hence

$$\frac{-f_p(0 \mid \mu)}{1 - f_p(0 \mid \mu)} \leq \rho \leq 1$$

$$\implies \frac{1}{1 - \exp(\mu)} \leq \rho \leq 1.$$

The number of zeros become less than expected in a Poisson distribution when:

$$\frac{-f_p(0 \mid \mu)}{1 - f_p(0 \mid \mu)} \leq \rho \leq 1.$$

This is known as zero-deflated Poisson distribution. The Poisson distribution is said to be truncated at zero when there are no excess zeros. The truncated Poisson distribution is obtained when:

$$\rho = \frac{-f_p(0 \mid \mu)}{1 - f_p(0 \mid \mu)}.$$

In this case the mixture distribution is given by:

$$f_{ZIP}\left(y \mid \rho, \mu\right) = \begin{cases} 0, & y = 0, \\ \frac{\mu^y}{[1-\exp(-\mu)]y!} \exp\left(-\mu\right), & y > 0. \end{cases} \tag{3.28}$$

The number of zeros become more than expected in a Poisson distribution $f_p\left(y \mid \mu\right)$ when $0 < \rho < 1$. In this case ρ represents the number of the excessive zeros. The mixture distribution $f_{ZIP}\left(y \mid \rho, \mu\right)$, is defined by the parameters ρ and μ. The mean and the variance of ZIP are given by $E(y) = (1 - \rho)$, and $var(y) = \mu(1-\rho)(1+\rho\mu)$, respectively. $E(y) < var(y)$ and $E(y) < \mu$ for $0 < \rho < 1$. In ZIP regression model, both ρ and μ must be modelled as the independent variables $\vec{X}_i$. This kind of a model uses log link to relate μ to the variables while the logit link is used to relate ρ to the variables. In ZIP regression model, the parameters ρ and μ may have covariates. In this case, we can use μ_i and v_i to represent two subsets of the covariates $\vec{X}_i$ connected to the parameters ρ and μ, respectively. The ZIP regression model is defined by $f_{ZIP}\left(y_i \mid \rho_i, \mu_i\right)$ with $logit(\rho_i) = \vec{u_i}^\top \vec{\beta_u}$ $1 < i < n$ $\forall \rho_i$ and $\log(\mu_i) = \vec{v_i}^\top \vec{\beta_v}$ $1 < i < n$ $\forall \mu_i$. The mean response and excess zeros are then modelled simultaneously in ZIP. Therefore, the model has one more link function for modelling the effect of independent variables on the excess zeros. Hence the excess zeros lead to biased estimates in ZIP regression model. The likelihood method is used for inference of ZIP regression model. Suppose the vector parameter $\vec{\theta} = \left(\vec{\beta_u}^\top, \vec{\beta_v}^\top\right)^\top$. Then, the distribution function for this ZIP model is given by:

$$f_{ZIP}\left(y_i \mid \vec{x_i}, \vec{\theta}\right) = \rho_i f_0\left(y_i \mid \rho_i\right) + \left(1 - \rho_i\right) f_p\left(y_i \mid \mu_i\right), \tag{3.29}$$

where:

$$logit(\rho_i) = \vec{\hat{u}_i}\vec{\beta_u}$$

and

$$\log(\mu_i) = \vec{v_i} be\vec{t}a_u, 1 < i < n.$$

Hence the log-likelihood function is given by:

$$\ell\left(\vec{\theta}\right) = \sum_{i=1}^{n} \log\left[\rho_i f_0\left(y_i \mid \rho_i\right) + \left(1-\rho_i\right) f_p\left(y_i \mid \mu_i\right)\right]. \tag{3.30}$$

3.3.5 Zero-Truncated Poisson and Zero-Truncated Negative Binomial regression models

Zero-Truncated Poisson (ZTP) and Zero-Truncated Negative Binomial (ZTNB) regression models are suitable for modelling zero-truncated data. These models are developed the same way as the Poisson and the Negative Binomial regression models. However, the ZTP and ZTNB models are modified to accommodate such zero-truncated count data. For a count variable say y to follow a truncated Poisson model, the variable should follow a Poisson distribution with 0 exclusive. The ZTP distribution is defined as:

$$f_{ztp}\left(y \mid \mu\right) = \frac{\lambda^Y \exp(-\lambda)}{y!\left[1-\exp\left(-\lambda\right)\right]}, \quad y > 0, y = 1, 2, ... \tag{3.31}$$

Hence, ZTP regression model is described by:

$$y_{i|\vec{X}_i} \sim ZTP(\mu_i), \log(\mu_i) = \vec{X_i^\top}\beta, 1 \leq i \leq n. \tag{3.32}$$

Inference for β under the ZTP model is also based on the MLE method. When we replace the ZTP distribution in (3.32) with the ZTNB distribution given by:

$$f_{ZTNB}\left(y \mid \mu, \alpha\right) = \frac{\Gamma(y+1/\alpha)}{y!(1/\alpha)\left[1-\left(\frac{1}{1+\alpha\mu}\right)^{1/\alpha}\right]}\left(\frac{\alpha\mu}{1+\alpha\mu}\right)^y\left(\frac{1}{1+\alpha\mu}\right)^{1/\alpha}, y = 1, . \tag{3.33}$$

we obtain the ZTNB model. This model also employs the MLE method for inferences about the model parameters.

3.3.6 Hurdle model

The Hurdle model is the model developed by modelling between-group difference and within-group difference. Within-group refers to the zero's that are at risk of being treated as positive count responses while between-group refers to the zero's that are not at risk of being treated as positive count responses in data modelling. The Hurdle model is specified as:

$$z_i \mid \vec{X}_i \sim Bernoulli(p_i), f(p_i) = \vec{X_i^\top}\vec{\beta},$$

$$y_i \mid z_i = 0, \vec{x_i} \sim ZTP(\mu_i), g(\mu_i) = \vec{X_i^\top}\vec{\beta}, \quad 1 \leq i \leq n, \tag{3.34}$$

where:

$z_i =$ between-group difference.

$y_i =$ positive count response.

Under the assumptions in (3.34), the likelihood function of this model is the product of the likelihood function of the binary component with $\vec{\alpha}$ as the only parameter vector and the likelihood for ZIP with $\vec{\beta}$ as the only parameter vector. Unlike the zero-inflated and zero-truncated models, the Hurdle model is conducted separately for each component. This model addresses a special case of the two-component mixture excluding sampling zero.

3.3.7 Maximum Likelihood Estimation method

Inference for parametric models is based on Maximum Likelihood estimation method. Suppose that $f\left(\vec{X}_i, \vec{\theta}\right)$ is the probability that $X_i = x$, where $\vec{\theta}$ is the parameter vector. Suppose that x_i represents a sample that is independently

and identically distributed for $1 \leq i \leq n$. The likelihood function for this sample is then given by:

$$L(\vec{\theta}) = \prod_{i=1}^{n} f\left(\vec{X}_i, \vec{\theta}\right).$$

For inference, the logarithm of $l(\theta)$, also known as the likelihood function is always used. This function is given by:

$$\ell\left(\vec{\theta}\right) = \sum_{i=1}^{n} \log f\left(\vec{X}_i, \vec{\theta}\right), \quad \vec{\theta} \in D, \tag{3.35}$$

where:

$D =$ the domain of $\vec{\theta}$.

The Maximum Likelihood Estimate (MLE) of $\vec{\theta}$ denoted by $\vec{\hat{\theta}_n}$ is obtained if the maximum of the likelihood function is obtained at an interior point $\vec{\hat{\theta}_n}$ of the domain of $\vec{\theta}$. Hence the derivative of the logarithm of $L(\vec{\theta})$ with respect to $\vec{\theta}$ must be 0 at $\vec{\hat{\theta}_n}$. Then $\vec{\hat{\theta}_n}$ is achieved by solving the score function denoted by:

$$W\left(\vec{\theta}\right) = \frac{\partial}{\partial\vec{\theta}}\ell(\theta) = \sum_{i=1}^{n} \frac{1}{f(\vec{X}_i, \vec{\theta})} \frac{\partial}{\partial\vec{\theta}} f(\vec{X}_i, \vec{\theta}) = 0. \tag{3.36}$$

The MLE is consistent and asymptotically normal, $\vec{\hat{\theta}_n} \sim N(\vec{\theta}, 1/n \sum)$, where $\sum = I^{-1}(\vec{\theta})$ and

$$I(\vec{\theta}) = -E\left[\frac{\partial^2}{\partial\vec{\theta}\partial\vec{\theta}^{\top}} \log f(\vec{x_i}, \vec{\theta})\right].$$

$I(\vec{\theta})$ is called the Fisher's Information matrix. The MLE is also considered to be asymptotically efficient. The hypothesis concerning the parameter vector $\vec{\beta}$ can be expressed as: $H_0 : c\vec{\beta} = a$ vs $H_1 : c\vec{\beta} \neq a$,

where:

$c =$ some known full rank $k \times 1$ matrix with $p(\geqslant k)$ denoting the dimension of $\vec{\beta}$.

$a =$ a known $k \times 1$ constant vector.

Both Wald and likelihood tests can be used to examine the general linear hy-

pothesis.

3.3.8 Canonical link function

The canonical link function is a natural link function to the family of distributions. For instance, in our Poisson regression model, we said that the link function is the log function. The general form of a link function $g(.)$, to a linear predictor η_i is defined as:

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}, \tag{3.37}$$

where:
$g(\mu_i) =$ the special case of $g(.)$ and
$\eta_i =$ the linear predictor.
As shown by (3.37), $g(.)$ works as a link between the RHS and LHS of the equation. Therefore, it is clear that the canonical link functions are derived directly from the density of a specified GLM. When the link functions are different, the interpretations of parameters also differ. Hence the selection of the link functions in the model is very critical. For canonical links, the derivative of the link is the same as the inverse of variance. The link function must be differentiable and monotonic. Hence (3.37) can also be written as:

$$\mu_i = g^{-1}(\eta_i) = g^{-1}\left(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}\right). \tag{3.38}$$

If we choose $g = h$, where $\theta = h(\mu)$, then

$$\theta_i = h(\mu_i) = h(h^{-1}(\eta_i)) = \eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}, \tag{3.39}$$

is the canonical link function, which ensures that the systematic component is modelling the parameters of interest. The canonical links also simplify the derivation of the MLE function, ensure that most of the assumptions of linear

regression model are not violated and also ensure that μ remains within the range of the response variable.

3.3.9 Exponential class

A one-parameter family of densities $f(.;\theta)$ that can be expressed in the form:

$$f(x,\theta) = a(\theta)(x)\exp\{c(\theta)d(x)\}\, for \quad all \quad x, all \quad \theta, \tag{3.40}$$

for a suitable choice of functions, $a(.), b(.), b(.)$ and $d(.)$ is said to belong to exponential family class. Hence a k-parameter family of densities $f(.;\vec{\theta})$ that can be expressed in the form:

$$f(x;\vec{\theta}) = a(\vec{\theta})b(x)\exp\left\{\sum_{i=1}^{k} c_i(\vec{\theta})d_i(x)\right\} for \quad all \quad x, all \quad \vec{\theta}, \tag{3.41}$$

for a suitable choice of functions, $a(.), b(.), b(.)$ and $d(.)$ is said to belong to the exponential family class.

3.4 Bayesian methods

The cornerstone of Bayesian framework is the theorem developed by Reverend Thomas Bayes. The framework combines the knowledge about the model parameters of a distribution of interest and the information about those parameters contained in the observed data. This combination is attained through the utilisation of Baye's theorem. The theorem results from an interconnection between the distribution's unconditional and conditional probability functions. Due to the uncertainties of the true values of parameters in the classical framework, the Bayesian framework considers the parameters as the random variables. Suppose that the parameter vector $\vec{\theta} = (\theta_1, ..., \theta_I)$ is a random variable and $\vec{Y} = y_1, ..., y_J$ denote the variable depending on $\vec{\theta}$. The parameter $\vec{\theta}$ is

unobservable. However, the inference about this parameter is based on Baye's theorem when given the observed data $\vec{Y}$. Bayesian universe is made up of all possible ordered pairs of the parameter vector $\vec{\theta}$ and all possible values of the observed random variable $\vec{Y}$. These pairs can be generally denoted by (θ_i, y_j) for $i = 1, ..., I$ and $j = 1, ..., J$. Each event $(\theta = \theta_1), ..., (\theta = \theta_J)$ partitions the Bayesian universe even though the events that have occurred remain unknown. Each event $(y = y_1), ..., (y = y_J)$ is always observed. Hence the events that have occurred are always known. Baye's theorem in this context is given by:

$$P(\vec{\theta} \mid \vec{Y}) = \frac{P(\vec{Y} \mid \vec{\theta}) * P(\theta)}{P(\vec{Y})}, \tag{3.42}$$

where:
$P(\vec{\theta} \mid \vec{Y}) =$ the posterior probability distribution that represents the knowledge about the parameters after inference.
$P(\vec{\theta}) =$ the prior probability distribution, which represents the prior knowledge about the parameter $\vec{\theta}$ before inference.
$P(\vec{Y} \mid \vec{\theta}) =$ the likelihood function that represents the relationship between the observed data and the parameter $\vec{\theta}$.
$P(\vec{Y}) =$ the marginal likelihood function which represents a normalisation factor.
We change the notations in (3.42) for simplicity. Suppose that $f(.)$ is a probability distribution encompassing the observable random variable $\vec{Y}$, and $g(.)$ be the probability distribution containing only the unobservable random variable parameter $\vec{\theta}$. Hence Baye's theorem in (3.42) will be given by:

$$f(\vec{\theta} \mid \vec{Y}) = \frac{f(\vec{Y} \mid \vec{\theta}) * g(\theta)}{f(\vec{Y})}, \tag{3.43}$$

where:
$f(\vec{Y})$ is the marginal distribution of a random variable $\vec{Y}$, given by:

$$f(\vec{Y}) = \int f(\vec{Y} \mid \vec{\theta}) * g(\vec{\theta}) d\theta. \tag{3.44}$$

Now, $f(\vec{Y})$ in (3.44) does not depend on the parameter $\vec{\theta}$ because it is obtained by averaging over all the possible values of $\vec{\theta}$. Hence (3.43) can be written as:

$$f(\vec{\theta} \mid \vec{Y}) \propto f(\vec{Y} \mid \vec{\theta}), \tag{3.45}$$

which indicates the posterior density of $\vec{\theta}$ up to some unknown constant. Therefore, in the Bayesian universe, each joint probability can be found by using the multiplication rule denoted by:

$$f(\theta_i \mid y_j) \propto g(\theta_i) * f(y_i \mid \theta_i). \tag{3.46}$$

3.4.1 Prior distributions

Prior distributions represent the prior knowledge about the parameters before the data are observed. This distribution serves as a key part of Bayesian inferential processes. Therefore, the strength of a posterior distribution depends on the strength of a prior distribution and the magnitude of the data available.

Informative priors

An informative prior distribution is a distribution for which the prior beliefs are significant in transforming the information contained in the data observed. Hence the conclusions about the model parameters based on the observed data and conclusions based on the prior distribution are different. Usually, the method used to select an informative prior distribution include the selection of a distribution for the unknown parameters and specify the parameters in

the selected distribution which reflect the prior knowledge about the unknown parameters. This implies that, the selected prior distribution must reflect the prior knowledge about the parameters. This is more important when selecting prior distributions for location parameters. Informative priors are usually referred to as conjugate priors. These priors only exist when the distribution that represent the prior knowledge belong to an exponential family class discussed in previous section by (3.40) and (3.41). In the conjugate family, the likelihood and the prior distribution functions are identical and the prior and the posterior distributions come from the same family of distributions. However, this is valid when observations are fixed and the parameters over all the possible values are different. Based on (3.22), we notice that the function $b(x)$ is only a scale factor. This implies that this function does not have any association with the spread of data. Therefore, $b(x)$ does not affect the shape of the prior distribution. For this reason, $b(x)$ can be absorbed into a constant of proportionality. The conjugate prior coming from one-dimensional exponential family of densities take the identical form as the likelihood function denoted by:

$$g(\theta) \propto a(\theta)^m e^{c(\theta)*n}, \tag{3.47}$$

where:
m and n are the constants that actuate the shape of the prior distribution.

Noninformative priors

Noninformative prior distributions are used when the prior knowledge about the parameters is ambiguous or not clear. Hence it is difficult to translate such knowledge into an informative prior. This kind of priors supply information that is clear and allow the information from the likelihood to be interpreted probabilistically. In most cases, noninformative priors are selected to be uniform probability distributions or the Jefrey's prior. The uniform probability

distribution selected as a prior probability distribution must be defined on the support of the parameter of interest. The noninformative priors have less impact on the posterior distribution compared to the observed data. These priors are also known as vague, diffuse, flat, weak, or reference prior distributions.

3.4.2 Bayesian linear regression models (BLMs)

This study covers univariate linear regression model. This model strives to explain variability in one dependent variable through the independent variables. This is achieved by assuming a linear relationship between these variables. Let us assume that the dependent variable $\vec{Y}$ has n observations. The model can be written as:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{k-1} x_{i,k-1} + \epsilon_i \quad i = 0, ..., n, \tag{3.48}$$

where:
y_i is the dependent variable, $i = 1, ..., n$,
x_{ik} is the independent variable, $p = 1, ...p - 1$,
β_0 is the regression intercept,
β_k is the regression coefficients, $p = 1, ..., p - 1$, and
ϵ_i is the regression disturbance.
The randomness of the relationship between the independent and dependent variables originates from the regression disturbance ϵ_i. The variability of the dependent variable explained by $x_k, k = 1, ..., k-1$ is represented by $\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{i,k-1}$. The distributional assumption about the source of randomness ϵ_i must be made. This is a way in which we can be able to describe the regression disturbance (ϵ_i). For simplicity, we assume that ϵ_i is independent and identically distributed with $N(0, \sigma^2)$. This implies that:

$$y_i \sim N(\mu_i, \sigma^2), \tag{3.49}$$

where:

$\mu_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{i,k-1}$. The expression in (3.49) can also be written in matrix form as:

$$\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}, \tag{3.50}$$

where: $\vec{Y}$ is an $n * 1$ vector denoted by:

$$\vec{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

$\vec{\beta}$ is a $k * 1$ vector denoted by:

$$\vec{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix},$$

$\mathbf{X}$ is an $n * k$ matrix denoted by:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,k-1} \\ 1 & x_{1,2} & \dots & x_{2,k-1} \\ \vdots & & & \\ 1 & x_{1,n} & \dots & x_{n,k-1} \end{pmatrix},$$

and $\vec{\epsilon}$ is a $n * 1$ vector denoted by:

$$\vec{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

The assumed distribution for the regression disturbance can also be written in a matrix form as:

$$\vec{\epsilon} \sim N(\vec{0}, \sigma^2 \mathbf{I}_n),$$

where: $\mathbf{I}_n$ is an $n * n$ identity matrix. Referring to the model presented in (3.50) we must estimate the parameters $\vec{\hat{\beta}}$ and σ^2. The likelihood function for the model is given by:

$$L(\beta_0, \beta_1, ...\beta_k, \sigma \mid Y, \vec{\mathbf{X}}) = (2\pi\sigma^2)^{-\frac{n}{2}} * \tag{3.51}$$

$$\exp\left\{ \frac{-1}{2\sigma^2} \sum_{i=1}^{k} (y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_{k-1} x_{i,k-1})^2 \right\}.$$

We can also write (3.51) in a matrix form as:

$$L(\vec{\beta}, \sigma \mid \vec{Y}\mathbf{X}) = (2\pi\sigma^2)^{-\frac{n}{2}} exp\left\{ \frac{-1}{2\sigma^2} \left(\vec{Y} - \mathbf{X}\vec{\beta}\right)^\top \left(\vec{Y} - \mathbf{X}\vec{\beta}\right) \right\}.$$

Equation (3.51) represents a multivariate normal distribution.

Estimation of BLMs

The Bayesian method of estimation accounts for the estimation risk of the parameters and also includes prior information to the model. We consider two prior scenarios:

- Noninformative prior

The joint noninformative conjugate prior for $\vec{\beta}$ and σ^2 is given by:

$$g(\vec{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}, \tag{3.52}$$

where:
the coefficients of regression can take any real value, $-\infty < \beta_p < \infty$ for $p = 1, ..., k$ and the disturbance variable is nonnegative, i.e $\sigma^2 > 0$. To obtain the posteriors of the model parameters, the likelihood in (3.51) and the prior in (3.52) are combined as:

$$f(\vec{\beta} \mid \mathbf{X}, \sigma^2) \sim N(\vec{\hat{\beta}}, (\mathbf{x}^\top \mathbf{x})^{-1}\sigma^2), \tag{3.53}$$

where:
$\vec{\hat{\beta}}$ is identical to the Maximum likelihood estimate of the classical method and $(\mathbf{x}^\top \mathbf{x})^{-1}\sigma^2$ is the covariance matrix of $\vec{\hat{\beta}}$. We note that the expression in (3.31), the posterior distribution of $\vec{\beta}$ conditional on σ^2 is recognised to follow a multivariate normal distribution. The posterior distribution of the regression disturbance (σ^2) follows an inverted chi-square distribution. This Posterior distribution is given by the inverse of:

$$p(\sigma^2 \mid \vec{Y}, \mathbf{X}) = \chi^2(n - k, \hat{\sigma}^2), \tag{3.54}$$

where:
σ^2 is identical to an estimate obtained through the MLE method. The marginal distribution of $\vec{\beta}$ is obtained by integrating (3.53) which results in:

$$p(\vec{\beta}, \sigma^2 \mid \vec{Y}, \mathbf{X}) = P(\vec{\beta} \mid \vec{Y}, \mathbf{X}, \sigma^2)P(\sigma^2 \mid \vec{Y}, \mathbf{X}), \tag{3.55}$$

with respect to σ^2 such that the marginal posterior distribution of $\vec{\beta}$ fol-

lows a multivariate student's t-distribution with a kernel given by:

$$P(\vec{\beta} \mid \vec{Y}, \mathbf{X}) \propto (n-k) + (\vec{\beta} - \hat{\vec{\beta}})^{\top} \frac{\mathbf{X}^{\top}\mathbf{X}}{\hat{\sigma}^2} (\vec{\beta}\hat{\beta})^{\frac{-n}{2}}. \tag{3.56}$$

The marginal distribution of the vector $\vec{\beta}$ becomes more heavily tailed when σ^2 is integrated. This shows the uncertainty about the true value of σ^2. Its variance increases with the term $\frac{v}{(v-2)}$ as follows:

$$\sum_{\vec{\beta}} \hat{\sigma}^2 (\mathbf{X}^{\top}\mathbf{X})^{-1} \frac{v}{v-2}, \tag{3.57}$$

where:

$v = n - k$ is the degrees of freedom for the distribution of $\vec{\beta}$. This is attained although the mean vector of $\vec{\beta}$ is not changed. Each standardised coefficient of the regression model follows a student's t-distribution with $n - k$ degrees of freedom as its marginal posterior distribution. This implies that if the parameter of interest is β_k only, then:

$$\left(\frac{\beta_k - \hat{\beta}_k}{(\eta_{k,k})^1/2} \mid \vec{Y}, \mathbf{X} \right) \sim t_{n-k}, \tag{3.58}$$

where:

$\eta_{k,k}$ is the k^{th} diagonal element of $\hat{\sigma}^2(\mathbf{X}^{\top}\mathbf{X})^{-1}$ and $\hat{\beta}_k$ is the estimate of β_k.

- Informative conjugate prior

  The natural conjugate prior reflects the available prior knowledge. They are more useful when the regression disturbance is assumed to follow the normal distribution. This helps in finding the convenient analytical posterior results. Hence we assume that $\vec{\beta}$ has a normal prior distribution given σ^2, which follows an inverted chi-square prior distribution. That is:

$$(\vec{a} \mid \sigma) \sim N(\vec{\beta_0}, \sigma\mathbf{A})$$

and

$$(\sigma^2) \sim Inv - \chi^2(v_0, c_0^2).$$

The parameters to be determined include β_0, v_0 and c_0^2. $\mathbf{A}$ is the scale matrix which is usually selected to be $\tau^{-1}(\mathbf{X}^\top\mathbf{X})^{-1}$. The purpose for this selection is to obtain a prior covariate that is identical to the one obtained via MLE of $\vec{\beta}$ up to a scaling constant. The degree of confidence that the mean of $\vec{\beta}$ is $\vec{\beta}_0$ can be adjusted through the distinction of the scale parameter τ. We fix the prior mean $\vec{\beta}_0$ at some default value to affirm it. However, that is not important if more specific prior information is available. The following prior simple data can be used to affirm the parameters of the inverted chi-square distribution:

$$v_0 = n_0 - k$$

and

$$c_{0^2} = \frac{1}{v_0}\left(\vec{Y}_0 - \mathbf{X}_0\vec{\hat{\beta}}_0\right)^\top\left(\vec{Y}_0 - \mathbf{X}_0\vec{\hat{\beta}}_0\right),$$

where the subscript 0 represents the prior data sample. In this case, the posterior distribution of the model parameter, $\vec{\beta}$ and σ^2 and prior distribution have the same form. The posterior distribution for the parameter vector $\vec{\beta}$ is given by:

$$p(\vec{\beta} \mid \vec{Y}, \mathbf{X}, \sigma^2) = N\left(\vec{\beta}^*, \Sigma_{\vec{\beta}}\right), \tag{3.59}$$

where:

$$\vec{\beta}^* = \left(\mathbf{A}^{-1} + \mathbf{X}^\top\mathbf{X}\right)^{-1}\left(\mathbf{A}^{-1}\vec{\beta}_0 + \bar{\mathbf{X}}\mathbf{X}\vec{\hat{\beta}}\right),$$

is the posterior mean, and

$$\Sigma_{\vec{\beta}} = \sigma^2\left(\mathbf{A}^{-1} + \mathbf{X}^\top\mathbf{X}\right)^{-1},$$

is the posterior variance. Although the posterior and the prior have the same form, their parameters are updated to show the observed data and prior beliefs. The mean of the posterior distribution is equal to the weighted mean of the prior distribution. The inverted chi-square distribution of σ^2 is given by:

$$p(\sigma^2 \mid \vec{Y}, \mathbf{X}) = Inv - \chi^2(\vec{v}^*, \mathbf{c}^{2*}), \tag{3.60}$$

and the parameters of the posterior distribution of σ^2 are given by:

$$\vec{v}^* = v_0 + n$$

and

$$\vec{v}^* \mathbf{c}^{2*} = (n-k)\vec{\hat{\sigma}}^2 + \left(\vec{\beta}_0 - \vec{\hat{\beta}}\right)^\top \mathbf{H} \left(\vec{\beta}_0 - \vec{\hat{\beta}}\right) + v_0 c_0^2,$$

where the parameters of the posterior distribution of $\vec{\beta}$ is obtained by integrating (3.59) which results in:

$$p(\vec{\beta} \mid \vec{Y}, \mathbf{X}, \sigma^2) \propto (\vec{v}^* + (\vec{\beta} - \vec{\beta}^*)^\top \mathbf{Q} (\vec{\beta} - \vec{\beta}^*))^{-1/2}, \tag{3.61}$$

where:

$$\mathbf{Q} = \frac{\left(\mathbf{A}^{-1} + \mathbf{X}^\top \mathbf{X}\right)}{\mathbf{C}^2}.$$

The marginal posterior distribution for each regression coefficient β_k is given by:

$$\left(\frac{\beta_k - \beta_k^*}{(q_{k,k})^{1/2}} \mid \vec{Y}, \mathbf{X}\right) \sim t_{v_0+n-k},$$

where:

$q_{k,k}$ is the k^{th} element of q^{-1}, and

β_k is the k^{th} component of $\vec{\beta}^*$.

Prediction

Our interest is to predict the response variable Y. Assume that we want to predict Y h steps ahead in time. We denote the future observations by $h * 1$ vector $\vec{\tilde{Y}} = (y_{T+1}, y_{T+2}, ..., y_{T+h})$. Suppose that the future observations of the explanatory variables are known and denoted by $\tilde{\mathbf{X}}$. Hence the predictive density is given by:

$$P(\vec{\tilde{Y}} \mid \vec{Y}, \tilde{\mathbf{X}}, \mathbf{X}) = \int \int P(\vec{\tilde{Y}} \mid \vec{\beta}, \sigma^2 \tilde{\mathbf{X}}) P(\vec{\beta}, \sigma^2 \mid \vec{Y}\mathbf{X}) d\vec{\beta}, \sigma^2, \tag{3.62}$$

where:
the joint posterior distribution of $\vec{\beta}$ and σ^2 is denoted by $p(\vec{\beta}, \sigma^2 \mid \vec{Y}, \mathbf{X})$. When using a noninformative prior, the predictive distribution is given by:

$$P(\vec{\tilde{Y}} \mid \vec{Y}, \tilde{\mathbf{X}}, \mathbf{X}) = t(n-k, \tilde{\mathbf{X}}, \vec{\hat{\beta}}, \mathbf{S}), \tag{3.63}$$

where:

$$\mathbf{S} = \hat{\sigma}^2(\mathbf{I}_p + \tilde{\mathbf{X}}(\mathbf{X}^\top \mathbf{X})^{-1} \tilde{\mathbf{X}}^\top)$$

and $\vec{\beta}$ is the posterior mean.
When using the informative conjugate prior, the predictive distribution is given by:

$$P(\vec{\beta}^* \mid \vec{Y}, \tilde{\mathbf{X}}, \mathbf{X}) = t(v_0 + n, \tilde{\mathbf{X}}, \vec{\beta}^*, \mathbf{V}), \tag{3.64}$$

where:

$$\mathbf{v} = c^{2*}(\mathbf{I}_p + \tilde{\mathbf{X}}(\mathbf{A}\mathbf{X}^\top \mathbf{X})^{-1} \tilde{\mathbf{X}}^\top),$$

and $\vec{\beta}^*$ is the posterior mean. Both the prior and posterior distribution in this case come from the same family of Normal distribution. We notice that the predictive distribution under noninformative and informative priors both follow a multivariate Student's t-distribution. Therefore, each component of $\vec{\tilde{Y}}$ follows

a univariate Student's t-distribution. Suppose we are interested in the component y_k. When using the noninformative prior, the predictive distribution is given by:

$$\left(\frac{\tilde{Y}_k - \tilde{\mathbf{X}}^K \hat{\beta}_k}{s_{k,k}^{1/2}} \right) \sim t_{n-k},$$

where $\tilde{\mathbf{X}}^K$ is the k^{th} row of $\tilde{\mathbf{X}}$, and $s_{k,k}$ is the k^{th} diagonal element of $\mathbf{S}$. When using the informative conjugate prior, the predictive distribution is given by:

$$\left(\frac{\tilde{Y}_k - \tilde{\mathbf{X}}^K \beta_k^*}{v_{k,k}^{1/2}} \right) \sim t_{v_0+n-k},$$

where:

$v_{k,k}$ is the k^{th} diagonal element of the scale matrix $\mathbf{V}$.

3.4.3 Computational approach to the Poisson regression model

The equations for the likelihood, prior and posterior distributions are not shown because the Markov Chain Monte Carlo (MCMC) algorithms are independent of the functional form of the posterior distribution. In the computational framework, the sample is drawn from the actual posterior distribution. However, it is easy to find the shape of the posterior distribution rather than its actual distribution. Metropolis-Hastings algorithm is the generalisation of other algorithms including Gabbs sampler algorithm. Hence this study is going to use Metropolis-Hastings algorithm with an independent candidate density to find the shape of the posterior distribution. The candidate density to be used must be very close to the posterior distribution. Therefore, many candidate densities will be accepted. Heavy tails of the candidate density as compared to the tails of the posterior distribution enables quicker movements within the parameter space. This helps in obtaining shorter burn-in and also use less thinning. The

starting point of the simulation process is the maximum likelihood vector $\vec{\beta}$ and the matched curvature matrix $\mathbf{v}$. The likelihood function is approximated by a multivariate normal distribution with the mean $\vec{\beta}$ and the covariance matrix $\mathbf{v}$. Informative prior for $\vec{\beta}$ can be used. This prior follows a multivariate Normal$[\vec{b_0}, \mathbf{v}_0]$ distribution. The approximate posterior distribution is given by:

$$g(\vec{\beta} \mid \vec{Y}) \propto g(\beta) f(\vec{Y} \mid \vec{\beta}). \tag{3.65}$$

Both the prior and the likelihood functions follow a multivariate normal distribution. This implies that the posterior distribution also follows a multivariate noramal distribution. Hence the updated constants will be given by:

$$\mathbf{v}^{-1} = \mathbf{v}_0^{-1} + \mathbf{v}^{-1} \longrightarrow \mathbf{v}_0^{-1} = 0,$$

and

$$\vec{b_1} = \mathbf{v}_1 \mathbf{v}_0^{-1} \vec{b_0} + \mathbf{v}_1 \mathbf{v}^{-1}.$$

A certain process is used to produce a candidate density from a multivariate Student's t-distribution. This distribution has a low degrees of freedom that matches the approximate posterior distribution. The process is outlined as:

- We use Cholesky decomposition to determine the lower triangular matrix $\mathbf{L}$ such that:

  $$\mathbf{L}\mathbf{L}^{\top} = \mathbf{V}_1.$$

  If we produce the matrix $\mathbf{Z}$, a multivariate normal$(\vec{0}, \mathbf{I})$ distribution of the right dimension by piling the independent normal$(0, i^2)$ variables, then the posterior distribution will be given by:

  $$\vec{w} = \vec{b_1} + \mathbf{L}\mathbf{V}. \tag{3.66}$$

  This posterior distribution follows a normal$(\vec{b_1}, \mathbf{v}_1)$ distribution.

- In this case we use the same transformation to get the candidate vector. However, we pile independent student's t-distribution random variables with low degrees of freedom to form a multivariate student's t-distribution denoted by t. The candidate vector is given by:

$$\vec{\beta} = \vec{b_1} + \mathbf{L}t.$$

The candidate vector $\vec{\beta}$ follows a multivariate student's t-distribution $t(\vec{b_1}, \mathbf{v}_1)$. The candidate density $\vec{\beta}$ matches the posterior close to the mode and also have heavier tails than the posterior. This process produces a random sample of candidates. This makes the movements within the parameter space to be quicker. All the candidates are accepted when the candidate density is identical to the true posterior distribution. In this case, burn-in and thinning are not required to obtain the random sample for inferences. However, the candidate density is usually not identical to the true posterior in practice although they are very similar. The similarity is attained because the shape of the candidate density matches the shape of the true posterior at the mode. Therefore, a good quantity of the candidates will be accepted. These candidates possess heavy tails. Hence movements within the parameter space will be very fast. Therefore, the burn-in will not be long and not much thinning will be required to obtain the approximate random sample from the posterior distribution. This sample is drawn from a true posterior distribution not the approximate posterior. Hence the credible intervals from this sample form a good representation of the claimed coverage probability provided that the sample size is also a good representative of the population it is drawn from.

The multivariate normal conjugate prior

The prior for the parameter vector follows a multivariate normal distribution, $N(\vec{b}_0, \mathbf{v}_1)$, where:

$$\vec{b}_0 = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix}$$

and

$$\mathbf{v}_0 = \begin{pmatrix} s_0^2 & 0 & \dots & 0 \\ 0 & s_1^2 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & s_p^2 \end{pmatrix}.$$

The predictor variables must be centred at their corresponding means. We rearrange the explanatory variables as $x_{ij} = x_{ij} - \bar{x}_{.j}$, where x_{ij} is the j^{th} value of the predictor for the i^{th} observation and $x_{.j}$ is the sample average of the j^{th} prediction. Hence it is easy to convert our prior knowledge about the mean of an average observation to the prior for the intercept $\vec{\beta}_0$ because the average observation affirms that all predictor values are equal to 0. This conversion is achieved by matching the percentiles.

Normal prior for intercept

Let us assume that we have 95% prior probability confidence that the mean of an average observation lies between l and u. We determine the normal(b_0, s_0^2) prior that matches this prior belief. This results in the following simultaneous equations:

$$l = e^{b_0 - 1.96 s_0}$$

and

$$u = e^{b_0 + 1.96 s_0}.$$

The solution for the simultaneous equations is given by:

$$b_0 = \frac{\log(l) + \log(u)}{2}$$

and

$$s_0 = \frac{\log(u) - \log(l)}{3.92}.$$

Normal prior for the slope

We now determine normal(b_j, s_j^2) prior for the slope coefficient β_j. This is attained by matching our prior belief about the ratio of the mean of the average observation for one unit increase in c_{ij} compared to the average observation. We have different procedures for the two cases presented:

- Case1: x_{ij} is a (0,1) indicator variable

  The average observations are in groups labelled 0 and 1. Two percentiles of the prior belief distribution of the ratio in group 1 to the mean of the average distribution in group 2 are matched. Let us assume we have 95% prior probability confidence that the ratio is between v and w. This results in the two equations:

  $$v = e^{b_j - 1.96 s_j}$$

  and

  $$w = e^{b_j + 1.96 s J}.$$

  The two equations solved simultaneously result with solutions:

  $$b_j = \frac{\log(v) + \log(w)}{2}$$

and

$$s_j = \frac{\log(w) - \log(v)}{3.92}.$$

- case2: x_j is a continuous variable

  The prior belief about the ratio of the mean of the average observations where x_j is increased by one standard deviation s_x to the mean of the average observations is matched. Let us assume that we have 95% prior probability confidence that the ratio is between v and w. This produces the two equations:

  $$v = e^{(b_j - 1.96 s_j) s_x}$$

  and

  $$w = e^{(b_j + 1.96 s_j) s_x},$$

  which gives the simultaneous solutions:

  $$b_j = \frac{\log(v) + \log(w)}{2 s_x},$$

  and

  $$s_j = \frac{\log(w) - \log(v)}{3.92 s_x}.$$

3.4.4 Computational Negative Binomial regression

Through the MCMC methods we use the Gibbs sampler for the NB regression model. This model is derived from the Poisson model to account for overdispersion which usually occurs in count data. Suppose the responses are independent. Then:

$$Y_i \sim NegBin(\lambda_i, r)$$

where:

$Y_i =$ the response variables for $i = 1, 2, ..., n$.

$r =$ the overdispersion parameter.

The expectation is modelled as:

$$\log(\lambda_i) = \mathbf{X}_i^\top \vec{\beta},$$

which implies that

$$\lambda_i = exp(\mathbf{X}_i^\top \vec{\beta}),$$

where:

$\mathbf{X} =$ the matrix of regressors.

$\vec{\beta} =$ the parameter vector.

The conditional likelihood of Y_i given w_i is defined as:

$$L(Y_i \mid r, \vec{\beta}, w_i) \propto exp\left\{k_i \mathbf{X}_i^\top \vec{\beta} - (\mathbf{X}_i^\top \vec{\beta})^2/2\right\} \tag{3.67}$$

$$\propto exp\left\{\frac{-w_i}{2}\left(\frac{y_i - r}{2w_i} - \mathbf{X}_i^\top \vec{\beta}\right)^2\right\},$$

where:

$k_i = \frac{y_i - r}{2}$. Exploiting property 1 of the poly-Gamma distribution, (3.67) can be written as:

$$l(Y_i \mid r, \vec{\beta}, w_i) = e^{k_i \eta_i} \int_0^\infty e^{-\psi_i \eta_i^2/2} p(\psi_i \mid r, Y_i, 0) d\psi_i, \tag{3.68}$$

where:

$\eta_i = \mathbf{X}_i^\top \vec{\beta}$. Suppose ψ_i is distributed according to PG$(Y_i + r, \eta_i)$, then following (Scott and Pillow, 2012), the conditional for $\vec{\beta}$ is given by:

$$p(\vec{\beta} \mid \vec{Y}^*, r, \vec{w}\vec{\psi}) \propto \pi(\vec{\beta}) exp\left[-1/2\left(z_i - \mathbf{X}^*\vec{\beta}\right)^\top \left(\mathbf{z} - \mathbf{X}^*\vec{\beta}\right)\Omega\right], \tag{3.69}$$

where:

$\vec{Y}^* =$ the $n * 1$ subvector of $\vec{Y}$ corresponding to w_i.

$n^* = \sum_{i=1}^{n} w_i$ is the number of individuals in risk class.
$\vec{\psi} =$ is a vector of length n^* with elements $z_i = \frac{y_i - r}{2\psi_i}$.
$\Omega diag(\psi_1, ..., \psi_n) =$ the $n * n$ precision matrix.
$\mathbf{X}^* = N^* * P$ matrix.
From (3.34), it is clear that $\vec{z}$ is normally distributed with mean $\vec{\eta} = \mathbf{X}^* \vec{\beta}$ and the diagonal covariance Ω^{-1}. Hence it is reasonable to assume a conditional Gaussian prior for $\vec{\beta}$ denoted by:

$$N_p\left(\vec{\beta}_0, \Sigma_0\right).$$

The conjugate prior full conditional distribution for $\vec{\beta}$ given $\vec{z}$ and Ω follows $N_p\left(\vec{\mu}, \Sigma\right)$, where:

$$\Sigma = \left(\Sigma_0^{-1} + \mathbf{X}^{*\top} \Omega \mathbf{X}^*\right)^{-1}$$

and

$$\vec{\mu} = \Sigma\left(\Sigma_0^{-1} + \mathbf{X}^{*\top} \Omega \vec{z}\right).$$

Therefore, given the current values for $\vec{\beta}, \vec{w}$ and r, the Gibbs sampler is given by:

- For w_i, draw ψ_i from its PG$(Y_i + r + \eta_i)$ distribution.
- For w_i, define $z_i = \frac{y_i - r}{2\psi_i}$.
- Update $\vec{\beta}$ from its N$(\vec{\mu}, \Sigma)$ distribution.
- Update r using a random-walk Metropolis-Hastings algorithm.

3.4.5 Goodness of fit

As discussed in the classical section, it is important to check how good the model is when analysing the data. However, in the Bayesian context, we are going to use the posterior predictive model checking to determine the goodness of fit. This method depends more on simulation based procedures. The method

is known as the posterior predictive distribution checking technique. This is a process by which the fit of the model to the observed data is assessed by drawing replicated data from the posterior predictive distribution. The distribution is given by:

$$p(\vec{y}^{rep} \mid \vec{y}) = \int p(\vec{y}^{rep} \mid \vec{\theta}) p(\vec{\theta} \mid \vec{y}) d\theta, \tag{3.70}$$

where:

$\vec{y} =$ the observed data vector.

$\vec{\theta} =$ the parameter vector.

$\vec{y}^{rep} =$ replicate data set vector.

Replicate data set is assumed conditionally independent from the observed data given the parameter. This data set is also assumed to be selected under the same conditions as the observed data. The posterior predictive distribution is obtained through the Markov Chain Monte Carlo (MCMC) methods. Potential successes of the model being fitted are observed when the replicated data drawn from the posterior predictive distribution is similar to the observed data. In this case, the probability that the replicated data could be more extreme than the observed data is measured by the p-value. Let us assume that $v(.)$ is a checking function that summarise data characteristics. The p-value of the posterior predictive distribution is given by:

$$p^B = P_r(v(\vec{y}^{rep})) \geqslant (v(\vec{y}) \mid \vec{y}), \tag{3.71}$$

which can be used to determine the variation between the observed data and data accumulated through simulations when the checking function is properly constructed.

3.4.6 Posterior inference

We recall from previous sections that the posterior distribution of a parameter vector $\vec{\theta}$ given the data observed $\vec{Y}$ is obtained through the application of Baye's theorem. We also recall that the posterior is the combination of the observed data and prior knowledge about the parameter of interest $\vec{\theta}$. Therefore, the posterior encompasses all the important information about the parameter $\vec{\theta}$. In Bayesian framework, inference is entirely based on the posterior distribution.

Bayesian point estimation

- Bayesian point estimation for quadratic loss function
  Let us consider the quadratic loss function:

$$L_2 = C(\hat{\theta} - \theta). \tag{3.72}$$

  Our interest is in finding the value of $\hat{\theta}$ that minimises the posterior mean square error given by:

$$E_{\theta|\vec{y}}[L_2] = \int c(\hat{\theta} - \theta)^2 g(\theta \mid \vec{y}) d\theta. \tag{3.73}$$

  Under the integral we differentiate and then we have:

$$\frac{d}{d\theta}\left\{E_{\theta|\vec{y}}[L_2]\right\} = \int 2c(\hat{\theta} - \theta) g(\theta \mid \vec{y}) d\theta,$$

  which when equated to zero to obtain the value of $\vec{\theta}$ that minimises $E_{\theta|\vec{y}}[L_2$. Hence we get:

$$\hat{\theta} \int g(\theta \mid \vec{y}) d\theta = \int \theta . g(\theta \mid \vec{y}) d\theta,$$

  from the properties of a proper density function, we notice that the point estimate for θ under quadratic loss function is the mean of the posterior

density, written as:

$$\hat{\theta} = E(\theta \mid \vec{y}) = \int \theta . g(\theta \mid \vec{y}) d\theta. \tag{3.74}$$

- Bayesian point estimation for a linear loss function
  We consider the linear loss function given by:

$$l_1 = \mid \hat{\theta} - \theta \mid . \tag{3.75}$$

The interest is in finding the value of $\hat{\theta}$ that minimises the posterior mean absolute deviation given by:

$$E_{\theta|\vec{y}}[L_1] = \int c \mid \hat{\theta} - \theta \mid^2 g(\theta \mid \vec{y}) d\theta \tag{3.76}$$

$$= \int_{-\infty}^{\theta} c(\hat{\theta} - \theta) g(\theta \mid \vec{y}) d\theta + \int_{\theta}^{\infty} c(\theta - \hat{\theta}) g(\theta \mid \vec{y}) d\theta.$$

By using differentiation under the integral signs we get:

$$\frac{d}{d\theta} \left\{ E_{\theta|\vec{y}}[L_1] \right\} = \int_{\theta}^{-\infty} c g(\theta \mid \vec{y}) d\theta - \int_{\infty}^{\theta} c g(\theta \mid \vec{y}) d\theta,$$

which after equating to zero we have:

$$\int_{-\infty}^{\hat{\theta}} g(\theta \mid \vec{y}) d\theta = \int_{\theta}^{\infty} g(\theta \mid \vec{y}) d\theta = 1/2,$$

which shows that $\hat{\theta}$ is the median of the posterior density function. This is noticed through the constant 1/2, which arose because the integrals are equal and sum to one.

Interval estimation

The point estimates for the central location of the posterior distribution are not informative enough when the ambiguities of the posterior distribution are significant. Therefore, the posterior $(1-\alpha)100\%$ credible interval $[a, b]$ is formulated to assess the degree of the ambiguities of the posterior distribution. The interval includes a specified probability that the random parameter θ is contained in the posterior. These intervals are called the credible intervals. The probability that the unknown parameter θ falls within a $(1-\alpha)100\%$ credible interval for the unknown parameter θ from the posterior is equivalent to finding an interval $(\theta_l.\theta_u)$ such that the probability of the posterior is given by:

$$(1-\alpha) = p(\theta_L < \theta_u) \tag{3.77}$$

$$= \int_{\theta_l}^{\theta_u} f(\theta \mid \vec{y}) d\theta.$$

The interval with the convergence probability required can be attained with no difficulties because there are many possible intervals meeting the requirement. However, the shortest interval (θ_l, θ_u) with the coverage probability required will have equal density values. In this case:

$$g(l \mid \vec{y}) = g(u \mid \vec{y}).$$

The interval (θ_l, θ_u) with equal tail areas is attained easier when we find θ_l and θ_u by:

$$\int_{-\infty}^{\theta_l} f(\theta \mid \vec{y}) d\theta = \alpha/2$$

and

$$\int_{\theta_u}^{\infty} f(\theta \mid \vec{y}) d\theta = \alpha/2.$$

One-sided hypothesis testing

A One-sided hypothesis test is conducted when we want to find out whether or not the treatment effect makes the parameter greater than the prior value it had for the standard treatment. We call the prior parameter value for the standard treatment the null value. When we want to find out whether θ is greater than the null value θ_0, we test the one-sided hypothesis by calculating the posterior probability of the null hypothesis given by:

$$\int_{-\infty}^{\theta_0} f(\theta \mid \vec{y}) d\theta, \tag{3.78}$$

using the posterior distribution. If the probability is less than the selected level of significance α, we can reject the null hypothesis in favor of the alternative hypothesis.

Two-sided hypothesis testing

We note that the continuous prior distribution results in a continuous posterior distribution. For this reason, the posterior probability of the point null hypothesis will always be equal to zero. Therefore, the two-sided hypothesis test cannot be attained by calculating the posterior probability of the null hypothesis as we did in one-sided hypothesis testing. Instead of constructing a two-sided hypothesis test, we calculate $(1-\alpha)100\%$ credible interval for the parameter θ. We then reject the null hypothesis if the null value θ does not lie in the credible interval. Hence we say the null value θ_0 is not a credible value.

Hypothesis comparison

The probability of a hypothesis can be computed to allow the comparison of true hypotheses. Suppose we want to compare the hypotheses:

$$H_0 : \theta \in \Theta_0$$

and

$$H_1 : \theta \in \Theta_1,$$

where Θ_0 and Θ_1 are the sets of all possible values for the unknown parameter θ. Hypotheses comparisons are based on θ's posterior distribution entirely as in point estimation and credible intervals. The posterior probability of the null hypothesis is determined by:

$$p(\theta \in \Theta_0 \mid \vec{y}) = \int_{\Theta_0} f(\theta \mid \vec{y}) d\theta, \tag{3.79}$$

and we compute the alternative hypothesis by:

$$p(\theta \in \Theta_1 \mid \vec{y}) = \int_{\Theta_1} f(\theta \mid \vec{y}) d\theta. \tag{3.80}$$

The posterior hypothesis probabilities show the prior knowledge and the observed data evidence about the parameter θ.

The posterior odds ratio

The posterior odds ratio is the odds ratio of the weighted likelihoods for the model parameters under the null hypothesis and under the alternative hypothesis, multiplied by the prior odds (Turner, 2008). This approach consists of summarising the two posterior hypotheses into a single value called the posterior odds ratio. The parameter uncertainties in this posterior odd ratio are taken account of because the weights are the prior parameter distributions. If

the prior probability of the null hypothesis is denoted by α, then the prior odds ratio is denoted by $(1-\alpha)$. The posterior odds are the prior odds updated with the information contained in the data. They are denoted by ρo and they are given by:

$$\rho o = \frac{\alpha}{1-\alpha}\frac{\int L(\theta \mid \vec{y}, H_0)g(\theta)d\theta}{\int L(\theta \mid \vec{y}, H_1)g(\theta)d\theta},$$

where:

$L(\theta \mid \vec{y}, H_0)$ is the likelihood function reflecting the restrictions imposed by the null hypothesis.

$L(\theta \mid \vec{y}, H_1)$ is the likelihood function reflecting the restrictions imposed by the alternative hypothesis. The prior odds are usually equated to one when no prior evidence is in favor or against the null hypothesis. We reject the null hypothesis when the value of the posterior odds in favor of the null hypothesis is low compared to the value of the posterior odds in favor of the alternative hypothesis. That is, we reject H_0 when

$$\rho o = \frac{\alpha}{1-\alpha}\frac{\int L(\theta \mid \vec{y}, H_0)g(\theta)d\theta}{\int L(\theta \mid \vec{y}, H_1)g(\theta)d\theta} < \frac{\alpha}{1-\alpha}\frac{\int L(\theta \mid \vec{y}, H_1)g(\theta)d\theta}{\int L(\theta \mid \vec{y}, H_0)g(\theta)d\theta}.$$

Then we can conclude that H_1 is more likely to be true as compared to H_0.

3.4.7 Bayesian numerical computation methods

Numerical computation methods were introduced for estimating complex models, especially in cases where the frequentist framework would need more efford and thus making the estimation methods more susceptible to errors. In the Bayesian framework, numerical computational methods are employed to generate samples from the posterior parameter distribution and predictive distributions in cases where analytical results cannot be obtained. Numerical computation methods are well known for increasing complex models manageability, although at high cost. These methods require that precise design of the sampling procedures be employed to attain more reliable posterior and predic-

tive inferences.

Monte Carlo integration

The Monte Carlo integration method can be utilised in cases where the posterior distribution of the parameter and the predictive distributions are verified as known distributions. This is usually in the case where conjugate prior distributions are used. For instance, suppose we are interested in estimating the posterior mean of a function $g(\vec{\theta})$. Let us denote the unknown parameter vector by $\vec{\theta}$ and the data observed by $\vec{y}$. The posterior mean of the function $g(\vec{\theta})$ is defined by:

$$Eg(\vec{\theta} \mid \vec{y}) = \int g(\vec{\theta}) p(\vec{\theta} \mid \vec{y}) d\theta, \tag{3.81}$$

where:
$p(\vec{\theta} \mid \vec{y})$ is the posterior distribution of the parameter $\vec{\theta}$. In cases where the equation above is impossible or difficult to evaluate analytically, we use approximation, which is obtained through the use of law of large numbers. Let us assume that we obtained a sample $\theta_1, ..., \theta_m$ from the posterior distribution $p(\vec{\theta} \mid \vec{y})$. When the sample size M approaches infinity, the quantity:

$$\hat{g_m}(\vec{\theta}) = \frac{1}{M} \sum_{m=1}^{M} g(\vec{\theta}_m),$$

converges to $E[g(\vec{\theta} \mid \vec{y})]$. This implies that, the approximation of the expected value of the function $g(\vec{\theta})$ becomes closer to the true expected value as the sample size from the posterior distribution is large. The basis for this approximation method is the Monte Carlo integration. The quantity $\hat{g_m}(\vec{\theta})$ denotes the Monte Carlo approximation, commonly known as the sample average. The quality of this approximation can be evaluated from the asymptotical statistics

results. The asymptotical variance of the Monte Carlo approximation $\hat{g_m}(\vec{\theta})$ is $\frac{\sigma^2}{M}$. The variance of the function $g(\vec{\theta})$ is σ^2 and can be estimated with the sample variance given by:

$$S_m^2 = \sqrt{1/M \sum_{m=1}^{M} [g(\vec{\theta}_m) - \hat{g_m}(\vec{\theta})]^2}.$$

Monte Carlo Standard Error (MCSE), the measure of numerical accuracy is given by:

$$MCSE = \sqrt{\frac{S_m^2}{M}}.$$

Monte Carlo approximation is outlined to be not the best method in practice because the estimators produced do not have the smallest approximation error and the posterior distributions that one often comes across in practice are not always in a known form. Therefore, the direct Monte Carlo approximation method is not always applicable. In such cases, the posterior and the predictive inference require the use of simulation algorithms that will be discussed in the next sections.

Markov chain

A stochastic process is a random process in discrete time. Any state of the process depends on the present state only and not on the past state. Markov chain is a special case of a stochastic process. In a stochastic process, the conditional probabilities at a time n given the states at all previous times $n-i, ..., 0$ depend only on one previous state at time $n-1$. Therefore,

$$p(x^{(2)} = x_2 \mid x^{(1)} = x_1, x^{(0)} = x_0) = p(x^{(2)} = x_2 \mid x^{(1)} = x_1),$$

$$p(x^{(3)} = x_3 \mid x^{(2)} = x_2, x^{(1)} = x_1, x^{(0)} = x_0) = p(x^{(3)} = x_3 \mid x^{(2)} = x_2),$$

and so forth. Markov chain's future evolution depends on the present state only. This process is said to possess the Markov property. The joint probability distribution of all states from time $0, ..., n$ can be constructed by:

$$p(x^{(1)} = x_1, x^{(0)} = x_0) = p(x^{(1)} = x_1 \mid x^{(0)} = x_0)p(x^{(0)} = x_0),$$

$$p(x^{(2)} = x_2, x^{(1)} = x_1, x^{(0)} = x_0) = p(x^{(2)} = x_2 \mid x^{(1)} = x_1)p(x^{(1)} = x_1 \mid x^{(0)} = x_0)p(x^{(0)} = x_0),$$

and so forth. If we are interested only in a state at time n, then we would add both sides over all possible values of $x_{n-1}, x_{n-2}, ..., x_0$. This results in the probability distribution at time n over all the possible states. The distribution is given by:

$$p(x^{(n)} = x_n) = \sum p(x^{(n)} = x_n \mid x^{(n-1)} = x_{n-1}) * p(x^{(n-1)} = x_{n-1}).$$

Then, denoting the random process by $\{x_n\}_{n=1}^{\infty}$, we can express the Markov chain by:

$$p(X_n = x_n \mid X_{n-1}x_{n-1}, X_{n-2} = x_{n-1,X_1=x_1}) = p(X_n = x_n \mid X_{n-1} = x_{n-1}).$$

In the context of posterior distribution, the state space (set of all possible states of a process) is the parameter space. For a Markov chain to converge to a long-run distribution, the properties such as irreducibility and ergodicity must be satisfied. However, the chains generated by MCMC satisfy these properties.

Markov Chain Monte Carlo sampling

As discussed in the previous section, we recall that the direct Monte Carlo intergration sampling from the posterior is ineffective when there is a large number of parameters or when the prior distribution is noninformative. Markov Chain Monte Carlo (MCMC) methods are then employed to generate samples

from the posterior distribution. In the context of MCMC methods, we set up a Markov chain that has the posterior distribution as its long-run distribution. This can be attained through the use of Metropolis-Hastings and Gibbs sampler algorithms. The MCMC methods are based on running the Markov chain long enough until it reaches the limiting (long-run) distribution. Therefore, any value taken after that initial run-in time approximates a random draw from the posterior distribution.

Metropolis-Hastings algorithm

Let us assume that we have parameter vector denoted by $\vec{\theta} = (\theta_1, \theta_2, ...\theta_p)$. Suppose that $q(\vec{\theta^\top}; \vec{\theta})$ is the candidate density when the chain is at $\vec{\theta}$ and let the posterior be denoted by $f(\vec{\theta} \mid \vec{y})$. The condition of reversibility can be expressed as:

$$f(\vec{\theta} \mid \vec{y})q(\vec{\theta}, \vec{\theta^\top}) = f(\vec{\theta^\top} \mid \vec{y})q(\vec{\theta^\top}, \vec{\theta}), \tag{3.82}$$

for all the possible states. Unfortunately, most chains cannot satisfy the reversibility condition for some $\vec{\theta}, \vec{\theta^\top}$. The probability of moving can be introduced to attain the balance. This moving probability is given by:

$$\alpha(\vec{\theta}, \vec{\theta^\top}) = min\left[1, \frac{f(\vec{\theta^\top} \mid \vec{y})q(\vec{\theta^\top}, \vec{\theta})}{f(\vec{\theta} \mid \vec{y})q(\vec{\theta}, \vec{\theta^\top})}\right]. \tag{3.83}$$

- Steps of Metropolis-Hasting algorithm:

1. Start at initial value $\theta^{(0)}$.
2. Do for $n = 1, ..., n$:

i. Draw $\theta^\top$ from $q(\theta^{(n-1)}, \theta^\top)$.

ii. Compute the probability $\alpha(\theta^{(n-1)}, \theta^\top)$.

iii. Draw u from $U(0,1)$.

iv. If $u < \alpha(\theta^{(n-1)}, \theta^{\top})$, then let $\theta^{(n)} = \theta^{\top}$, else let $\theta^{(n)} = \theta^{(n-1)}$.

Metropolis-Hastings with a random-walk candidate density

The candidate density is drawn from a distribution that is symmetric and centred at the current value. Using the parameter vector $\vec{\theta} = (\theta_1, \theta_2, ..., \theta_p)$, the random-walk candidate density is given by:

$$q(\vec{\theta}, \vec{\theta^{\top}}) = q_1(\theta_1^{\top} - \theta_1, ..., \theta_p^{\top} - \theta_p), \tag{3.84}$$

where:
for each argument the function $q(.)$ is symmetric about 0. Hence the candidate density can be written as:

$$q(\vec{\theta}, \vec{\theta^{\top}}) = q_1(\vec{\theta^{\top}} - \vec{\theta}), \tag{3.85}$$

where:
$q(.)$ is the vector function that is symmetric about the vector $\vec{0}$. Therefore, the acceptance probability for a random-walk candidate density is given by:

$$\alpha(\vec{\theta}, \vec{\theta^{\top}}) = min\left[1, \frac{f(\vec{\theta^{\top}} \mid \vec{y}) q(\vec{\theta^{\top}}, \vec{\theta})}{f(\vec{\theta} \mid \vec{y}) q(\vec{\theta}, \vec{\theta^{\top}})}\right] \tag{3.86}$$

$$= min\left[1, \frac{f(\vec{\theta^{\top}} \mid \vec{y})}{f(\vec{\theta} \mid \vec{y})}\right].$$

This implies that a candidate $\vec{\theta^{\top}}$ with a bigger value of the target density than the target density of the current value $\vec{\theta}$ has a 100% chance of being accepted. In this case the Markov chain will always move uphill. However, when a candidate $\vec{\theta^{\top}}$ has a lower value of the target density than the target density of the current value $\vec{\theta}$, $\vec{\theta^{\top}}$ will only be accepted with a probability identical to the

proportion of the target density value to the current value. Nevertheless, there is a chance that the chain will move downhill. This makes it possible for a random-walk candidate density to move around the entire parameter space.

Metropolis-Hastings with an independent candidate density

Since an independent candidate density is used, the density for which the candidate is drawn from does not depend on the current value. The independent candidate distribution is then defined as:

$$q(\vec{\theta}, \vec{\theta^{\top}}) = q_2(\vec{\theta^{\top}}). \tag{3.87}$$

The acceptance probability for the Markov chain using an independent candidate density is given by:

$$\alpha(\vec{\theta}, \vec{\theta^{\top}}) = min\left[1, \frac{f(\vec{\theta^{\top}} \mid \vec{y})q(\vec{\theta^{\top}}, \vec{\theta})}{f(\vec{\theta} \mid \vec{y})q(\vec{\theta}, \vec{\theta^{\top}})}\right] \tag{3.88}$$

$$= min\left[1, \frac{f(\vec{\theta^{\top}} \mid \vec{y})}{f(\vec{\theta} \mid \vec{y})} * \frac{q_2(\vec{\theta})}{q_2(\vec{\theta^{\top}})}\right].$$

Blockwise Metropolis-Hasting algorithm

The parameter vector is divided into blocks:

$$\vec{\theta} = \vec{\theta}_1, \vec{\theta}_2, ..., \vec{\theta}_J,$$

where:
$\vec{\theta}_j$ is a block of parameters. Suppose $\vec{\theta}_{-j}$ represents all the parameters that are not in block j. It is not easy to find a single overall kernel that converges to the joint density of the posterior compared to finding the conditional kernel for one block of parameter at a time that converges to its respective conditional density of the posterior. Hence Hastings (1970) advised that the Metropolis-Hastings

algorithm be applied successively to one block of parameters $\vec{\theta}_j$ after the other conditional to the knowledge of all the parameter values not belonging to the block.

- Steps of blockwise Metropolis-Hasting algorithm:

1. Start at a point in parameter space $\vec{\theta}_1^{(0)}, \vec{\theta}_2^{(0)}, ..., \vec{\theta}_j^{(0)}$.

2. For $n = 1, ..., N$, for $j = 1, ..., J$:

i. draw candidate from

$$q\left(\vec{\theta}_j^{(n-1)}, \vec{\theta}_j^{\top} \mid \vec{\theta}_1^{(n)}, ..., \vec{\theta}_{j-1}^{(n)}, \vec{\theta}_{j+1}^{(n-1)}, ..., \vec{\theta}_J^{(n-1)}\right).$$

ii. Calculate the acceptance probability:

$$\alpha\left(\vec{\theta}_j^{(n-1)}, \vec{\theta}_j^{\top(n)}, \vec{\theta}_1^{(n)}, ..., \vec{\theta}_{j-1}^{(n)}, \vec{\theta}_{j+1}^{(n-1)}, ..., \vec{\theta}_J^{(n-1)}\right).$$

iii. Draw u from $U(0,1)$ if $u < \vec{\theta}_j^{(n-1)}, \vec{\theta}_j^{\top}$ then let $\vec{\theta}_j^{(n)} = \vec{\theta}_j^{\top}$, else let $\vec{\theta}_j^{(n)} = \vec{\theta}_j^{(n-1)}$.

When given all the parameters $\vec{\theta}_{-j}$ and the observed data $\vec{y}$, the candidate density for the parameter block $\vec{\theta}_j$ must dominate the true conditional density in the tails. That is:

$$q(\vec{\theta}_j, \vec{\theta}_j^{\top} \mid \vec{\theta}_{-j}) > f(\vec{\theta}_j \mid \vec{\theta}_{-j}, \vec{y}).$$

At each step with the block in turn, the candidate $\vec{\theta}_j$ is drawn from the candidate density. Then, acceptance probability is calculated. The block of parameters is either moved to the candidate $\vec{\theta}_j^{\top}$, or kept at the current value $\vec{\theta}_j$. This depends on whether or not a random draw from $U(0,1)$ random variable is larger than the acceptance probability.

3.4.8 Gibbs sampling

Gibbs sampling is based on the blockwise Metropolis-Hastings algorithm discussed in previous sections. Let us assume that at each step for each block of parameters given others, we use the true conditional density in place of the candidate density. This implies that:

$$q(\vec{\theta}_j, \vec{\theta}_j^{\top} \mid \vec{\theta}_{-j}) > f(\vec{\theta}_j \mid \vec{\theta}_{-j}, \vec{y}).$$

Therefore, at step n for block $\vec{\theta}_j$, the acceptance probability is given by:

$$\alpha\left(\vec{\theta}_j^{(n-1)}, \vec{\theta}_j^{\top(n)}, \vec{\theta}_1^{(n)}, ..., \vec{\theta}_{j-1}^{(n)}, \vec{\theta}_{j+1}^{(n-1)}, ..., \vec{\theta}_J^{(n-1)}\right)$$

$$= min\left[1, \frac{f(\vec{\theta}_j^{\top} \mid \vec{\theta}_{-j}, \vec{y}) q(\vec{\theta}_j^{\top} \mid \vec{\theta}_{-j})}{f(\vec{\theta}_j \mid \vec{\theta}_{-j}, \vec{y}) q(\vec{\theta}_j, \vec{\theta}_j \mid \vec{\theta}_{-j})}\right] = 1.$$

The candidate will always be accepted at each step. Gibbs sampling is a special case of blockwise Metropolis-Hastings algorithm. This is a case where each candidate block is drawn from its true conditional density provided that all the other blocks given are at their recently drawn values.

3.4.9 Convergence diagnostics

The credibility of a posterior inference based simulation algorithms depend on the convergence of Markov Chain. When the Markov chain has reached convergence, the simulated sample is indeed drawn from the desired posterior distribution. The important goal of posterior distribution is to generate a Markov chain which moves around the entire parameter space easily. In some posterior simulation, Markov chain is unable to move well around the parameter space or even get trapped for long periods of time. This kind of Markov chain is undesirable and can be generated when autocorrelations between successive parameter draws are high and their decay is slow. Convergence is not

prevented by high correlation. However, it leads to delay for convergence to be reached. The influence of Markov chain starting point reduces as the number of iterations increase and in the end the starting point cannot be traced. Part of the chain simulation is discarded to minimise the influence of the chain's initial state. The portion discarded is referred to as the burn-in fraction. Hence the remaining portion of the chain's simulation are used in posterior inference. The size of the burn-in fractions is determined by the mixing speed of Markov chain. Fast mixing Markov chains tend to forget their origin after several iterations. Therefore, half of the iterations discarded are needed for chains displaying high serial correlation of the draws. There are methods for assessing and monitoring convergence. This method depends on examining the behavior of different quantities characterising the posterior distribution. The Markov chain has reached the stationary distribution when the quantities characterising the posterior distribution exhibit very divergent values at various points of the simulation sequence. We discuss two convergence monitoring methods.

Cumsum convergence monitoring

This is a simple monitoring tool where the trace plot of the standardised posterior means that are taken as the number of iterations are visually inspected. Convergence is represented by a stable dynamic. The statistic is defined as:

$$cs_{i,m} = 1/m \frac{\sum_{j=i}^{m} (\theta_i^{(j)} - \hat{\theta}_i)}{\hat{\sigma}_i}, \tag{3.89}$$

where:

$m =$ after-burn-in number of simulations.

$\hat{\theta} =$ the posterior mean.

$\hat{\sigma} =$ the posterior standard deviation of θ_i.

The Markov chain is said to converge when the value of the statistic in (3.89) approaches zero.

Parallel chains convergence monitoring

This method is based on running several independent chains in parallel. These chains must have different starting values. When the chain produces outputs that are very similar, convergence is reached. The similarity of the outputs is determined by how close the average variance of the after-burn-in simulations for a certain chain is to the variance of the posterior means across the chains. Let us assume we are interested in a parameter θ and R parallel chains are run. $\theta^{(i,r)}$ denote the $i^{th}(i = 1, ..., M)$ simulation of θ from the $r^{th}(r = 1, ..., R)$ chain. The mean within-sequence variation is estimated by:

$$W = 1/R \sum_{r=1}^{R} \hat{\sigma}_r^2,$$

where:

$$\hat{\sigma}_r^2 = \frac{\sum_{i=1}^{m} \left(\theta^{(i,r)} - \hat{\theta}^{(r)}\right)^2}{M-1}$$

and

$$\hat{\theta}^{(r)} = \frac{\sum_{i=1}^{m} \theta^{(i,r)}}{M}.$$

The between-sequence variation is estimated by:

$$B = \frac{M}{R-1} \sum_{r=1}^{R} (\hat{\theta}^{(r)} - \hat{\theta})^2,$$

where:

$$\hat{\theta} = 1/R \sum_{r=1}^{R} \hat{\theta}^{(r)}.$$

The posterior variance θ is estimated as a weighted average of W and B:

$$var(\hat{\theta}) = \frac{M-1}{M} * W + \frac{1}{M} * B,$$

where we suppress the condition on the observed data $\vec{y}$ notationally. The chain starts from far-apart initial values of the parameter. Hence the within-sequence variation will be smaller than between-sequence variation before convergence. When convergence has been reached, $var(\hat{\theta})$ is close to W. Hence, the statistic is given by:

$$Q = \frac{var(\hat{\theta})}{W}.$$

When the value of Q is close to 1, convergence is reached. When the value of Q is much greater than 1, the chain must continue to run until it reaches convergence.

3.4.10 Summary of the chapter

This chapter provides the background for the area of the study, which is the Limpopo province of South Africa. Malarial count data of interest are described in detail. Poisson distribution is pointed out to be suitable for modelling count data. However, the assumptions of Poisson distribution are not always satisfied. Therefore, the Poisson models developed under the dissatisfied (mean is not equal to variance) Poisson assumptions are overdispersed. Several methods that can account for such overdispersion are discussed. The methods are based on the development of the refined models such as NB, ZIP, ZINB, ZTP, ZTNB and Hurdle models. The classical method of estimation, the MLE is also discussed. The Bayesian framework and its background are outlined. These included the Bayesian linear regression model. The Bayesian framework is easier to execute through the employment of computational Bayesian approaches. Therefore, computational approaches to Poisson regression model are discussed. Furthermore, the model to account for an overdispersion, the computational NB model is also discussed. The Bayesian method of estimation, MCMC is discussed. To assess how good, the developed model fit the data, both the classical and the Bayesian paradigms make use of goodness of

fit tests. The comparisons between the Bayesian and the classical framework will be based on the errors executed by each method. These errors are known as the standard errors in the classical framework and as naive standard errors in the Bayesian framework.

Chapter 4

Results and discussion

4.1 Introduction

This chapter presents the results and discussions of the study. The results are obtained through the application of the methodology outlined in Chapter 3 and the R codes used to obtain the results are outlined in the appendix. These results are intended to fulfill the following research objectives, which are to:

i. Model malaria incidence given rainfall, temperature, normalised vegetation index, elevation and time in quarters from 2014 to 2015 across the various districts of the Limpopo province.

ii. Identify the effect of environmental factors which require more attention towards malaria control and prevention in Limpopo province.

iii. Examine the behavioral changes (trends) in overall malaria incidence in Limpopo province.

iv. Identify districts that are more susceptible to malaria incidence.

This chapter includes descriptive statistics. This is attained through the construction of bar charts for the categorical variables, histogram and scatter plots for the continuous variables to complement the distributions of malaria incidence. Various models are developed and compared. The best estimation method is also identified.

4.2 Exploratory data analysis

As outlined in previous chapters, malaria data used in this study is sourced from Malaria Institute based in Tzaneen, Limpopo. Population data were attained from StatsSA, while the data for the environmental factors were obtained from Ecoverb. The variables are described in Table 4.1 and the data summary is presented in Table 4.2.

Table 4.1: Variable description

Variable	Description	Data set code
Malaria	The number of malaria cases.	mal
Population	The population size.	pop
Districts	The five districts of Limpopo Province: Capricorn, Mopani, Sekhukhune, Vhembe, and Waterberg.	dist
Years	The years for which the data were collected: 2014 and 2015.	dyear
Elevation	The elevation above sea level measured in meters.	ele
Rainfall	The rainfall measured in millimeters.	rain
NDVI	The difference between near-infrared reflected by the vegetation and red light which is absorbed by vegetation, it ranges from -1 to 1.	ndvi
Temperature during the day	The maximum temperature in degrees Celsius.	td
Temperature during the night	The minimum temperature in degrees Celsius.	tn

Table 4.2: Summary descriptive statistics of the variables

Statistics	Ele	Tn	Td	NDVI	Rain	Mal
Min	19.69	4.925	20.99	0.2060	0.00	0.00
1^{st} Quartile	182.43	13.171	28.72	0.3247	0.5165	1.00
Median	242.20	16.7128	32.44	0.4035	23.9065	3.00
Mean	325.27	16.251	31.66	0.4133	32.5911	23.95
3^{rd} Quartile	491.18	19.558	34.64	0.4973	47.0245	12.25
Maximum	822.24	25.601	40.24	0.6670	159.2010	4820
Std deviation	216.09	4.450	4.15	0.1056	38.4265	61.83

Figure 4.1 presents a histogram used to display the incidence of malaria across the entire province of Limpopo. This aids in visual determination of the distribution followed by the response variable, malaria counts.

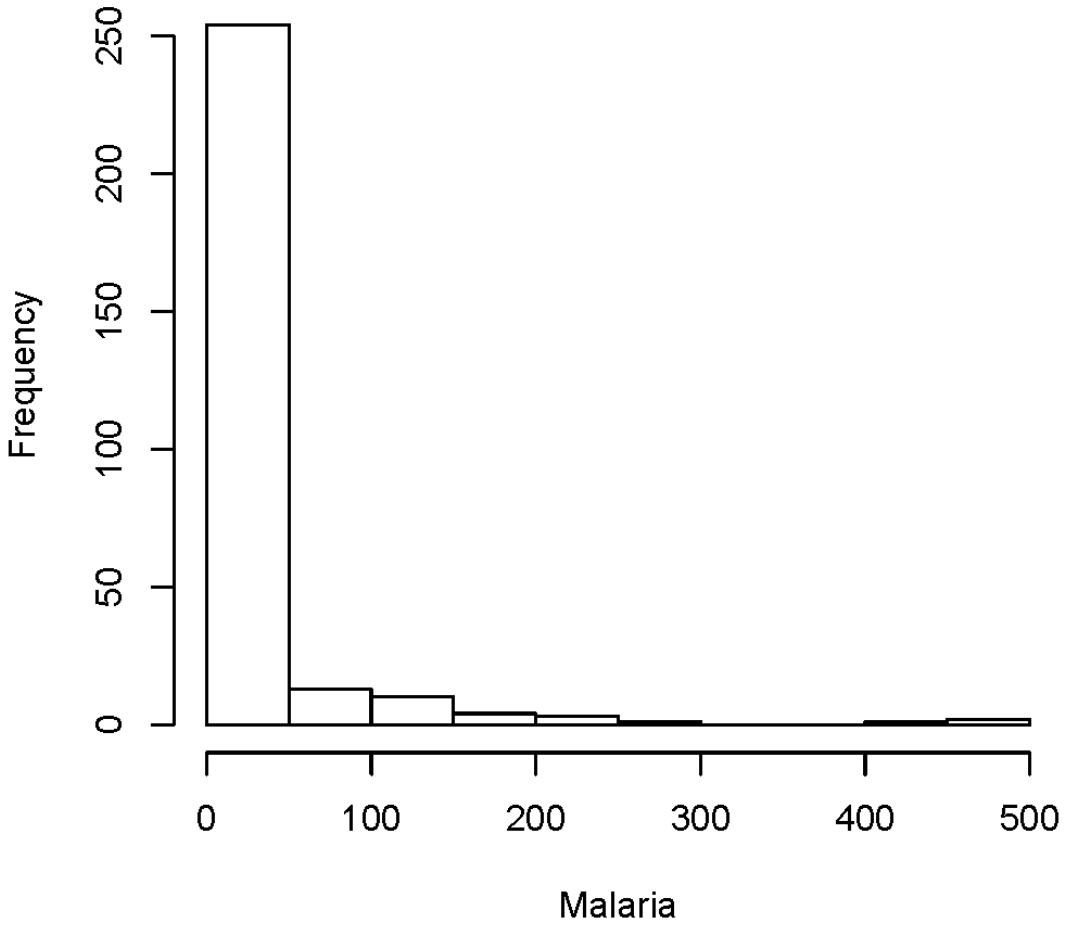

Figure 4.1: Histogram for malaria distribution

The histogram depicted in Figure 4.1 is skewed to the right. It takes a lopsided mound shape with its tail going off to the right. The shape of this histogram is similar to the shape of a Poisson distribution (Consul and Jain, 1973).

The scatter plots are used to depict the relationship between malaria counts and each predictor variable. These are displayed in Figure 4.2 to Figure 4.8.

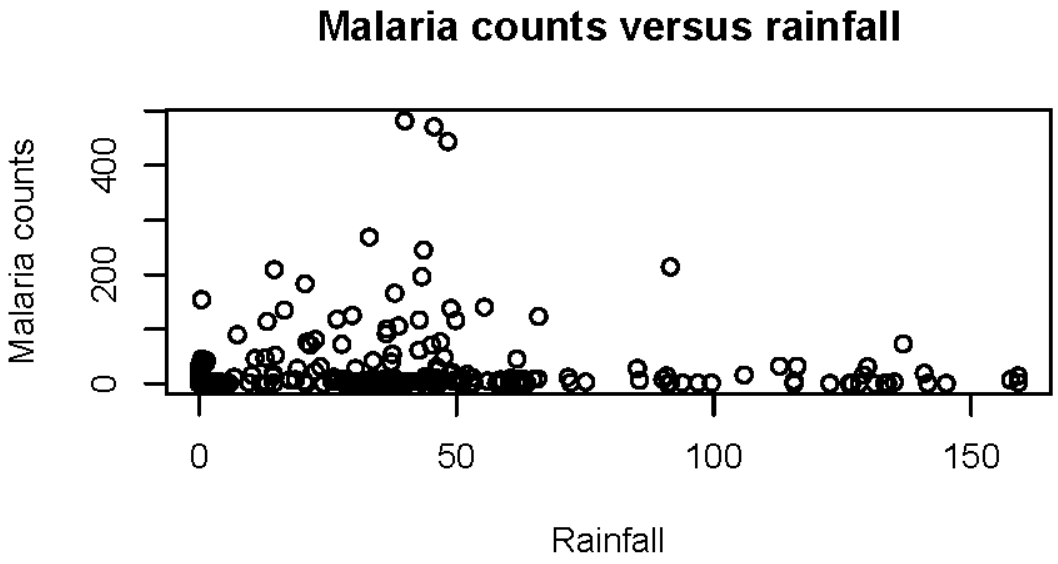

Figure 4.2: The distribution of malaria incidence with respect to rainfall

Figure 4.2 shows the relationship between malaria incidence and rainfall. This relationship appears to be non-linear. Malaria incidence rate is shown to be high between 0 and 50 millimeters of rainfall. However, the rate of malaria incidence decreases with an increasing amount of rainfall. Hence Figure 4.2 depicts a negative relationship between malaria and rainfall.

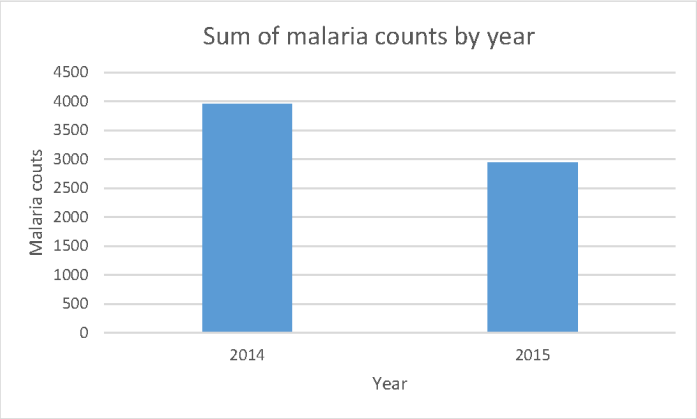

Figure 4.3: The distribution of malaria incidence in 2014 and 2015

According to Figure 4.3, the transmission rate of malaria was high in 2014 than in 2015. This may be due to various effects of enviromental factors as they may differ in each year and it may also indicate the success of malaria control, prevention and elimination methods that are used currently in Limpopo.

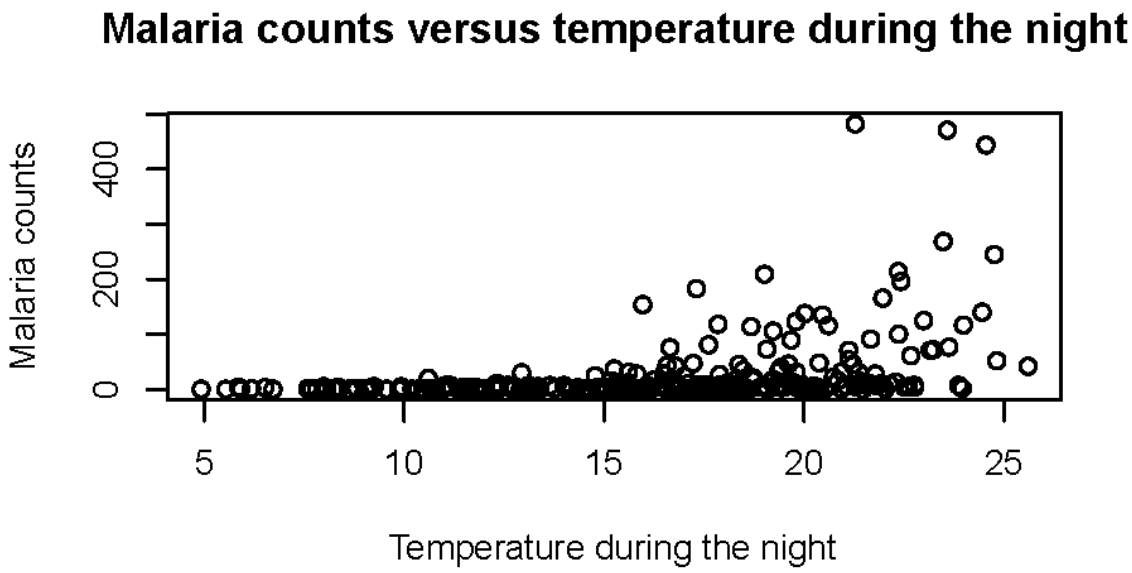

Figure 4.4: The distribution of malaria incidence over temperature during the night

Malaria cases are increasing gradually with an increasing temperature during the night (Figure 4.4). This implies that the relationship between malaria incidence and temperature at night is positive. The transmission of malaria is displayed to be high between $15°C$ and $25°C$, whereas it is low between $5°C$ and $14°C$.

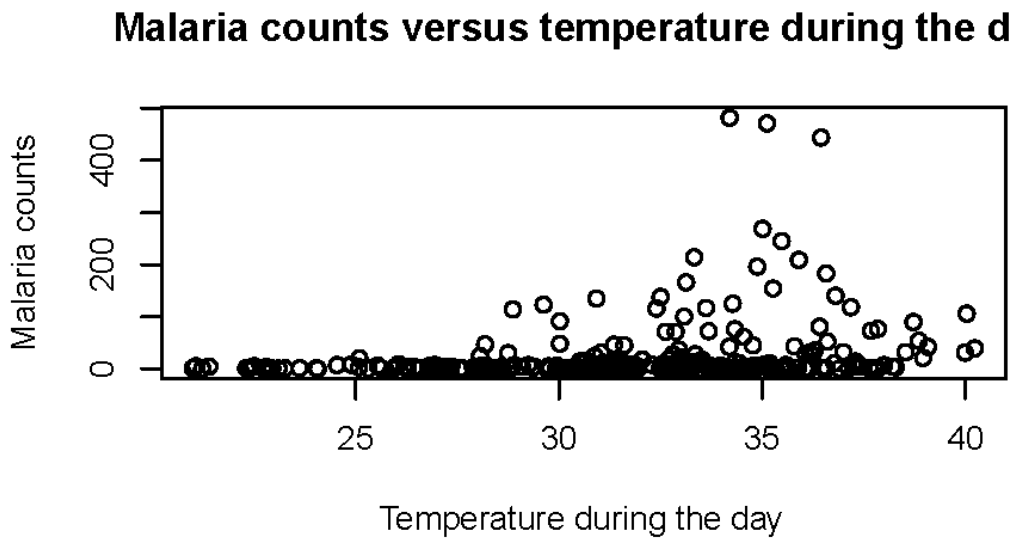

Figure 4.5: The distribution of malaria incidence over temperature during the day

The rate of malaria incidence remains constantly low at a temperature below $27°C$ during the day (Figure 4.5). However, it increases rapidly when the temperature is between $28°C$ and $37°C$. The association between the temperature during the day and malaria incidence is shown to be non-linear. Malaria transmission peaks between the temperature of $34°C$ and $37°C$ and slowly goes down as the temperature increases beyond $37°C$.

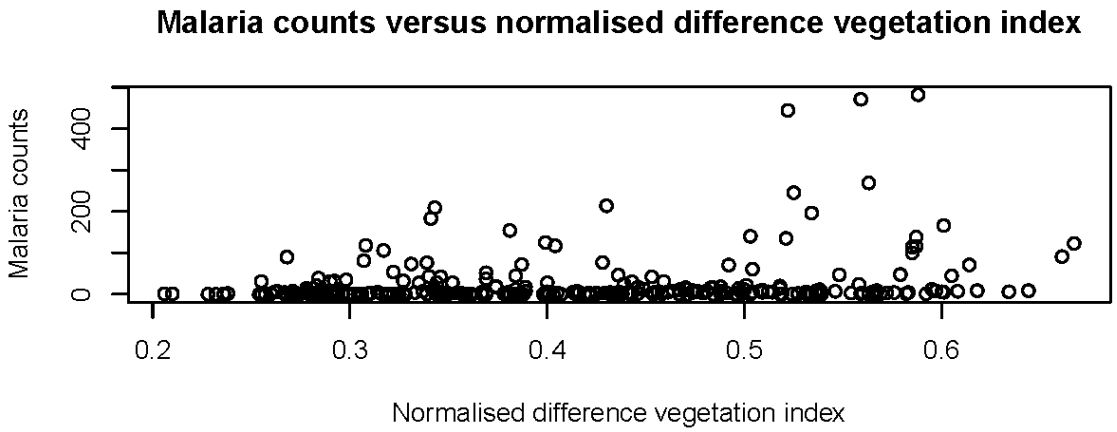

Figure 4.6: The distribution of malaria incidence over NDVI

Figure 4.6 depicts that malaria cases are not increasing at a constant rate. There are a lot of fluctuations in the distribution. However, as the vegetation becomes healthier (more green), the risk of malaria transmission also becomes high. Therefore, the association between malaria incidence and NDVI is shown to be positive.

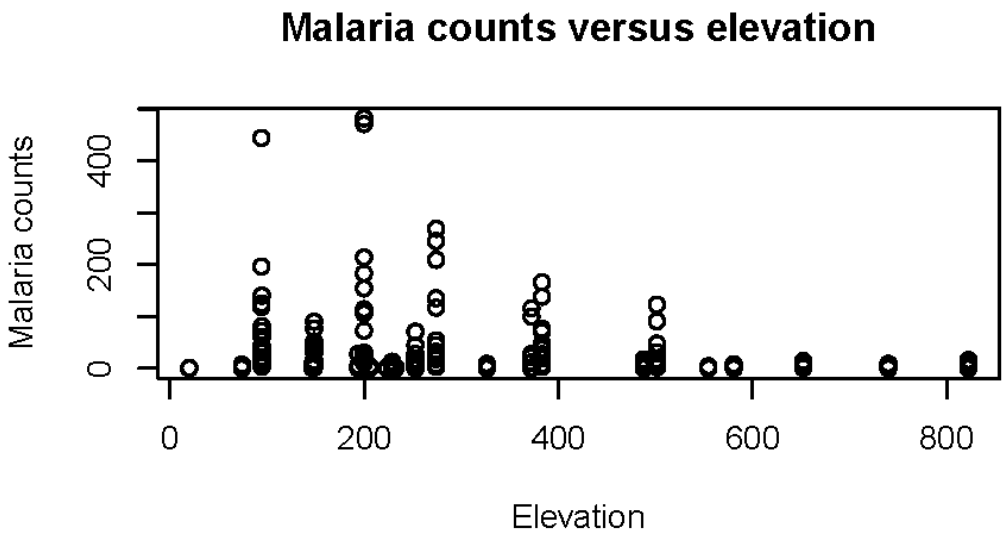

Figure 4.7: The distribution of malaria incidence over elevation

There are many cases of malaria between 100 and 400 meters above the sea level compared to the number of malaria cases between 400 and 800 meters above the sea level (Figure 4.7). This implies that the risk of malaria transmission decreases with an increasing elevation. Hence the relationship between malaria incidence and elevation is non-positive.

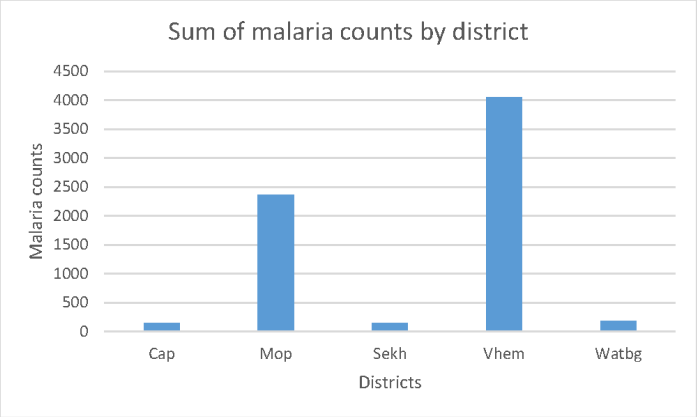

Figure 4.8: The distribution of malaria incidence across the districts of Limpopo

As shown in Figure 4.8, Vhembe district is depicted to have the highest rate of malaria incidence, followed by Mopani district as compared to all the other districts. Capricorn district has the lowest rate of malaria incidence. The high rate of malaria incidence in Mopani and Vhembe districts could be attributed to the high temperatures in the two districts.

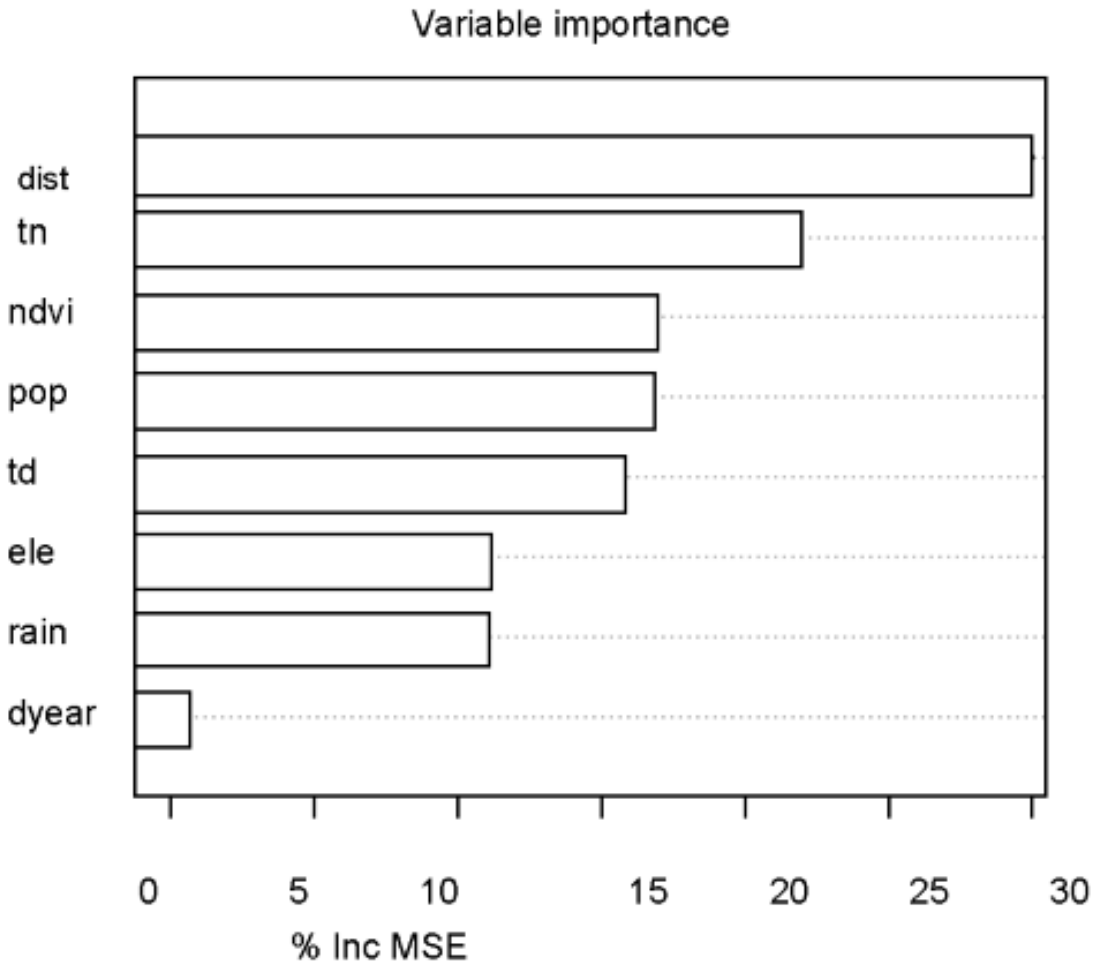

Figure 4.9: Variable importance graph in MSE percentages

Figure 4.9 shows that, the top 3 most imporatant variables are districts, temperature during the night and NDIV. Furthermore, the graph shows that the bottom 3 of the least important variables include year, rain and elavevation above sea level.

4.3 Model fitting

4.3.1 Classical methods

The development of the models is based on the distribution of the response variable, malaria counts. As discussed earlier in Chapter 3, most common models suitable for count data include: Poisson regression model and Negative Binomial regression model. As shown in Figure 4.1, the shape of malaria counts distribution is similar to the shape of a Poisson distribution. Therefore, three Poisson regression models are developed. Tables 4.3-4.6 present the summaries of the three Poisson models and their parameters. These models are used to perform the goodness of fit tests.

Table 4.3: Poisson model encompassing all the explanatory variables

Coefficient	Estimate	Std.Error	P-value	95% confidence interval
Intercept	-17.850	0.288	<0.001	[-18.415 : -17.288]***
Rain	-0.007	0.000	<0.001	[-0.009 : -0.007]***
Cap(Ref)	- - - -	- - - -	- - - -	- - - -
Mop	2.213	0.089	<0.001	[2.048 : 2.385] ***
Sekh	0.537	0.122	<0.001	[0.297 : 0.777]***
Vhem	2.458	0.084	<0.001	[2.297 : 2.627]***
Watbg	0.489	0.111	<0.001	[0.272 : 0.709]***
2014(Ref)	- - - -	- - - -	- - - -	- - - -
2015	-0.170	0.029	<0.001	[-0.227 : -0.114]***
tn	0.276	0.007	<0.001	[0.263 : 0.290]***
td	0.064	0.008	<0.001	[0.049 : 0.080]***
ele	-0.001	0.000	<0.001	[-0.001 : -0.001]***
ndvi	0.521	0.228	<0.001	[0.074 : 0.969]***
Key:	p <0.001 '***'	0.01 '**'	0.05 '*'	$\geqslant$0.05 ' '

The p-values for all the covariates as displayed in Table 4.3 are less than the level of significance, 0.05. This suggests evidence against the null hypothesis of no relationship between the covariates and malaria incidence. Rainfall and elevation estimate values are negative, suggesting that they have a negative

relationship with malaria incidence. Table 4.3 also depicts a positive relationship between temperature, NDVI and malaria incidence.

Table 4.4: Deviance and AIC for Poisson model in Table 4.3

Deviance	**Estimate**	**Df**
Null deviance	18541.3	287
Residual deviance	3532.6	277
AIC	4421.4	

According to Section 3.3.2 of Chapter 3, if the ratio of the deviance statistic and its degrees of freedom is significantly larger than 1, then there is an evidence of lack of fit in the model developed. Using Table 4.4, the ratio of the deviance statistic and its degrees of freedom is given by:

$$\frac{Residual\ \ deviance}{Df} = \frac{3532.6}{277} = 12.753.$$

The resulting value is significantly larger than 1. Hence there is evidence of lack of fit for the model presented in Table 4.3.

Table 4.5: Poisson model with exclusion of the district explanatory variable

Coefficient	**Estimate**	**Std.Error**	**P-value**	**95% confidence interval**
Intercept	-80.900	0.273	$<$0.001	[-18.6309 : -17.5624]***
Rain	-0.006	0.001	$<$0.001	[-0.007 : -0.005]***
2014(Ref)	- - - -	- - - -	- - - -	- - - -
2015	-0.025	0.027	$<$0.001	[-0.078 : 0.028]***
tn	0.239	0.007	$<$0.001	[0.226 : 0.252]***
td	0.125	0.008	$<$0.001	[0.110 : 0.014]***
ele	-0.002	0.000	$<$0.001	[-0.003 : -0.002]***
ndvi	3.049	0.212	$<$0.001	[2.635 : 3.465]***
Key:	$p <$0.001 '***'	0.01 '**'	0.05 '*'	$\geqslant$0.05 ' '

The p-values for all the covariates displayed in Table 4.5 are extremely significant (the p-values are very close to zero) at 5% level of significance. This implies that there is a relationship between the covariates and malaria incidence. The

estimate values for the regression coefficients of the covariates rainfall and elevation are negative, revealing that each of these covariates exhibit a negative relationship with malaria incidence. Table 4.5 also depicts a positive relationship between malaria incidence and each of the covariates temperature and NDVI.

Table 4.6: Deviance and AIC for Poisson model in Table 4.3

Deviance	**Estimate**	**Df**
Null deviance	18541.3	287
Residual deviance	6387.5	281
AIC	7268.3	

Based on Table 4.6, the ratio of the deviance statistic and its degrees of freedom is given by:

$$\frac{Residual\ \ deviance}{Df} = \frac{6387.5}{281} = 22.7313.$$

The resulting value is significantly larger than 1. Hence there is evidence of lack of fit in the model presented by Table 4.5.

Table 4.7: Poisson model with exclusion of the NDVI explanatory variable

Coefficient	**Estimate**	**Std.Error**	**P-value**	**95% confidence interval**
Intercept	-17.410	0.213	<0.001	[-17.829 : -16.994]***
Rain	-0.008	0.000	<0.001	[-0.009 : -0.007]***
Cap(Ref)	- - - -	- - - -	- - - -	- - - -
Mop	2.235	0.085	<0.001	[2.071 : 2.406] ***
Sekh	0.507	0.121	<0.001	[0.268 : 0.745]***
Vhem	2.483	0.083	<0.001	[2.323: 2.651]***
Watbg	0.510	0.111	<0.001	[0.294 : 0.728]***
2014(Ref)	- - - -	- - - -	- - - -	- - - -
2015	0.051	0.025	<0.001	[-0.252 :-0.155]***
tn	0.288	0.004	<0.001	[0.280: 0.297]***
td	0.051	0.005	<0.001	[0.041 : 0.061]***
ele	-0.001	0.000	<0.001	[-0.001 : -0.001]***
Key:	p <0.001 '***'	0.01 '**'	0.05 '*'	⩾0.05 ' '

The summary of the model presented in Table 4.7 shows that the p-values for

all the covariates are close to 0. This suggests evidence against the null hypothesis of no association between the covariates and malaria incidence. Rainfall and elevation estimate values are negative. This suggests that they have a negative relationship with malaria incidence. Table 4.7 also reveals a positive relationship between temperature and malaria incidence.

Table 4.8: Deviance and AIC for Poisson model in Table 4.3

Deviance	**Estimate**	**Df**
Null deviance	18541.3	287
Residual deviance	3537.8	278
AIC	4424.6	

Based on Table 4.8, the ratio of the deviance statistic and its degrees of freedom is given by:

$$\frac{Residual\ \ deviance}{Df} = \frac{3537.8}{278} = 12.7259.$$

The resulting value is significantly larger than 1, which implies lack of fit for the model in Table 4.7.

Model selection

Tables 4.4, 4.6 and 4.8 include the Akaike Information Criterion (AIC), which represents the measure of the information loss during the model fitting. According to Mazerolle (2006), the model with the lowest AIC is considered to be the best model. The AIC shown in Table 4.4 for the model in Table 4.3 is 4421.4, the AIC shown in Table 4.6 for the model in Table 4.5 is 7268.3 and the AIC shown in Table 4.8 for the model in Table 4.7 is 4424.6. Therefore, the model with the smallest AIC is the model presented in Table 4.3. Hence the Poisson model including all the covariates is the best model compared to the other developed Poisson models which excluded some of the covariates.

Detection of overdispersion

According to Hilbe (2011), Pearson's chi-square is considered to be the best method in detecting an overdispersion in Poisson models. If the ratio of the residual deviance and the degree of freedom is significantly larger than 1, then the probability that the developed model is overdispersed is high. Based on the best selected Poisson model, the ratio of the residual deviance and the degrees of freedom is 12.753. This implies that the probability that the selected Poisson model is overdispersed is very high. To validate that the Poisson model selected may be overdispersed, we check if the response variable satisfies the Poisson assumption of an equality between the mean and the variance. Table 4.2 shows that the mean for the response variable is 23.95 and through the use of R software, the variance is found to be 3822.5. Therefore, the condition of equal mean and variance for a Poisson distribution is violated. We can then conclude that the selected Poisson model presented by Table 4.3 is overdispersed.

The correction of overdispersion

Several methods for the correction of overdispersion are discussed in Chapter 3. However, the method suitable for correcting the overdispersed Poisson model presented in Table 4.3 is the development of a Negative Binomial (NB) model. It is outlined in Chapter 3 that the variance of the NB distribution is always larger than its mean and the overdispersion is naturally accounted for in the NB models. Hence the NB model is a suitable solution for our overdispersed Poisson model. Table 4.9 presents the NB regression model developed to correct the overdispersed selected model presented in Table 4.3.

Table 4.9: NB model encompassing all the explanatory variables

Coefficient	Estimate	Std.Error	P-value	95% confidence interval
Intercept	-15.330	1.038	<0.001	[-17.483 : -13.186]***
Rain	-0.005	0.002	0.012	[-0.0095 : -0.001]*
Cap(Ref)	- - - -	- - - -	- - - -	- - - -
Mop	2.215	0.209	<0.001	[1.789 : 2.643] ***
Sekh	0.415	0.258	0.107	[-0.097 : 0.925]
Vhem	2.848	0.211	<0.001	[2.412 : 3.285]***
Watbg	0.871	0.229	<0.001	[0.395 : 1.348]***
2014(Ref)	- - - -	- - - -	- - - -	- - - -
2015	0.206	0.125	0.100	[-0.040 : 0.452]
tn	0.254	0.033	<0.001	[0.181 : 0.326]***
td	0.000	0.033	0.999	[-0.071: 0.070]
ele	0.000	0.000	0.336	[-0.001: 0.001]
ndvi	-0.477	1.014	0.638	[-2.555 : 1.613]
Key:	p <0.001 '***'	0.01 '**'	0.05 '*'	⩾0.05 ' '

Table 4.9 presents the NB model developed to correct the overdispersed Poisson model presented in Table 4.3. According to Table 4.9, the p-value of the covariate rain is 0.012, which is less than 0.05. This implies that the covariate rain is significant at 5% level of significance. Hence, there is a relationship between rainfall and malaria incidence. The coefficient estimate of rain is negative. This implies that the relationship between rainfall and malaria incidence is negative. That is, malaria transmission rate increases with a decreasing amount of rainfall.

The p-values for Mopani, Vhembe and Waterberg are less than 0.001. These p-values suggest that there is a certain pattern of malaria transmission between these districts and Capricorn district (the reference category). The coefficient estimates for Mopani, Vhembe and Waterberg are positive. These estimates entail that if malaria incidence increases in each of these districts, then it also increases in Capricorn district (the reference category). We use the odds ratio, e^{β}, to find the precise pattern of malaria incidence amongst the districts. The

odds ratio in this case is the ratio of the odds of the reference category (Capricorn) and each of the districts among Mopani, Vhembe and Waterberg. If there is an increase in malaria incidence, the increase is e^{β} times more in Mopani, Vhembe and Waterberg than in Capricorn district. The Greek letter β in the odds ratio e^{β}, represents the regression coefficient. Table 4.9 provides evidence that malaria incidence increases by $e^{2.215} \approx 9$ times in Mopani, $e^{2.848} \approx 17$ times in Vhembe and $e^{0.871} \approx 2$ times in Waterberg than in Capricorn district.

A unit increase in temperature during the night increases the incidence of malaria by $e^{0.254} \approx 2$ times. There is no evidence of an existing association between malaria incidence and the covariates, temperature during the day, elevation and NDVI according to Table 4.9.

Table 4.10: Deviance and AIC for NB model in Table 4.9

Deviance	**Estimate**	**Df**
Null deviance	1232.31	287
Residual deviance	317.87	277
AIC	1680.9	

Based on Table 4.10, the ratio of the deviance statistic and its degrees of freedom is given by:

$$\frac{Residual\ \ deviance}{Df} = \frac{317.87}{277} = 1.148.$$

The resulting value is significantly close to 1 compared to the ratio of the deviance statistic and its degrees of freedom for the Poisson model presented in Table 4.3. Hence there is an evidence of best fit in the model presented by Table 4.9.

4.3.2 Bayesian methods

Bayesian inference is based on the posterior distribution. This distribution is obtained through the utilisation of Baye's theorem. The posterior distribu-

tion is the combination of the observed data and prior knowledge about the parameters of interest. Bayesian framework is easy to work with through its computation methods. Hence this study applies the Bayesian computational method, the MCMC to generate samples from the posterior distribution. The sample is truly drawn from the posterior distribution if its Markov chain has reached convergence. The Bayesian framework is outlined in detail in Chapter 3 of this study. Inferences for the classical framework are based on the NB model presented in Table 4.9 since the selected Poisson model presented in Table 4.3 was proved to be overdispersed. Therefore, we are only focusing on the MCMC method of estimation using the NB model.

Convergence of the Markov chains

The Figures 4.9-4.12 display the Markov chains trace plots and the kernel density plots. Each trace plot presents the values of the sampled parameter (y-axis) at each step of the Markov chain (x-axis). These steps are commonly known as the iterations. The kernel density plots estimate the posterior marginal distributions for each parameter with the parameter values on the x-axis and the density on the y-axis.

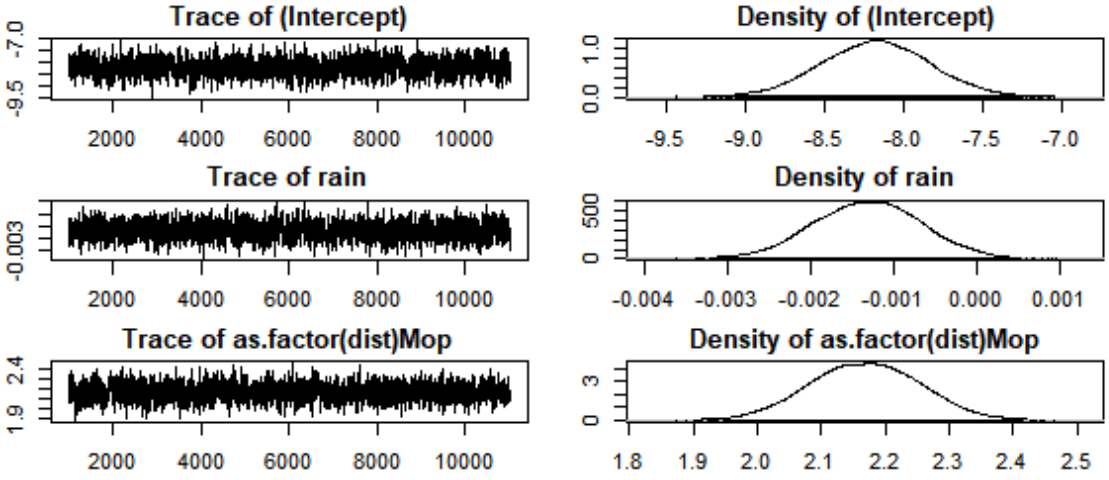

Figure 4.10: The trace plots and marginal densities for the intercept and the coefficients of covariates rain and Mopani district

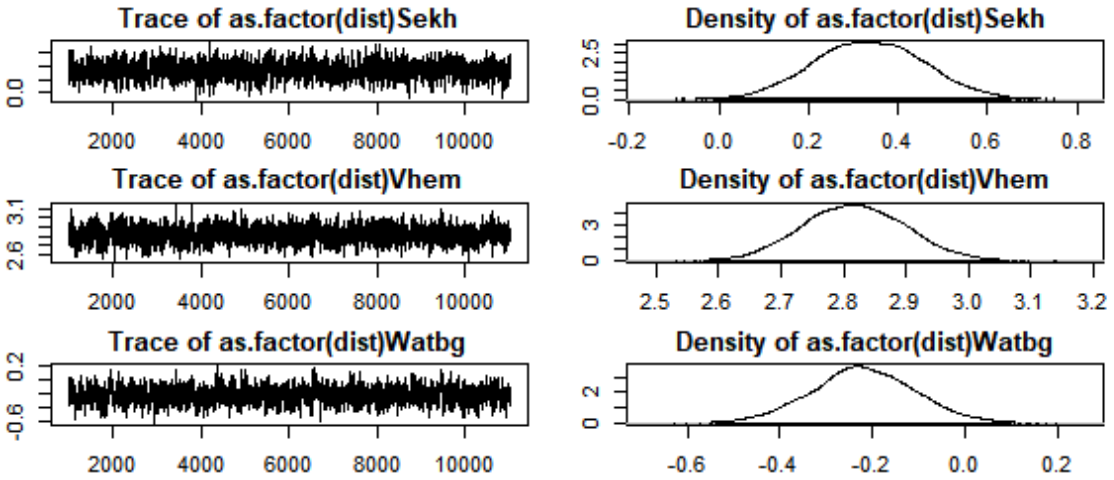

Figure 4.11: The trace plots and marginal densities for Sekhukhune, Vhembe and Waterberg districts

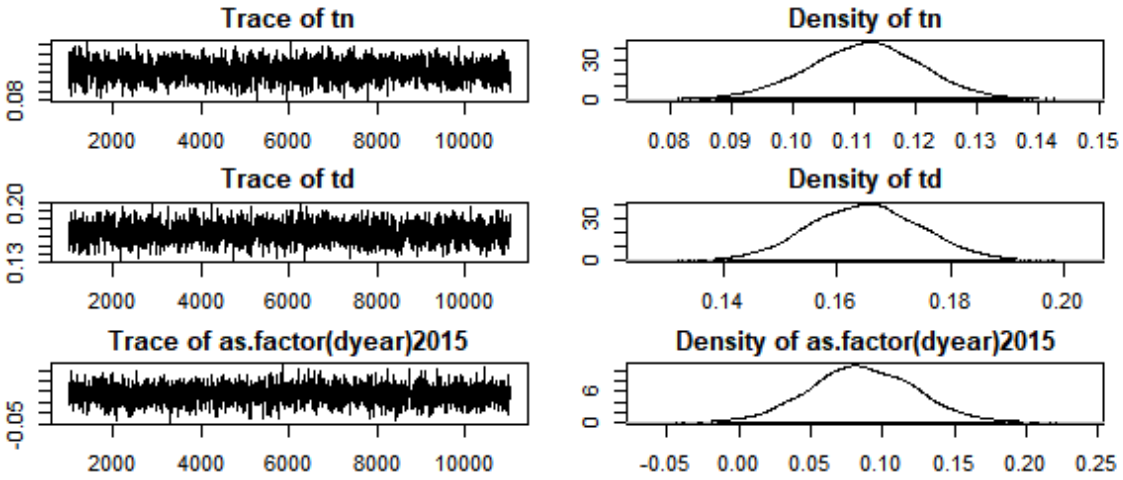

Figure 4.12: The trace plots and marginal densities for the year 2015 and temperature

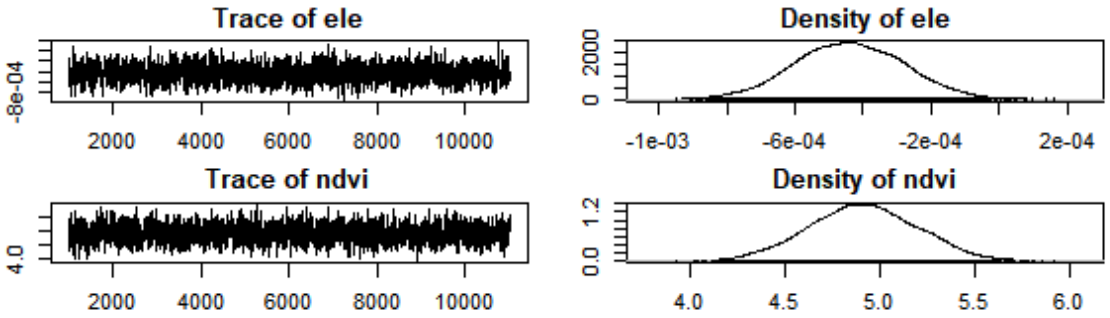

Figure 4.13: The trace plots and marginal densities for the covariates elevation and NDVI

All the trace plots presented in Figures 4.9-4.12 show that the Markov chains appear to have reached their stationary distributions. The mean for each Markov chain has stabilised and appear to be constant over the plots. The stationarity of the Markov chain is also affirmed by the bell-shape of the marginal distributions, showing that the posterior distribution is normalised. This implies that the sample is truly drawn from the posterior distribution. Hence the inferences based on this sample are reliable and informed.

The posterior inferences

A 100(1-α) % equal-tail interval corresponds to the 100(α/2) and 100(1-α/2) percentiles of the posterior distribution. However, this study will utilise the 95% highest posterior density (HPD) credible intervals to make the posterior inferences. The HPD credible interval is an interval in which most of the distribution lies and it is the one with the smallest width among all the credible intervals of the posterior distribution (Institute, 2014). Table 4.11 presents the NB regression model developed through the employment of the MCMC estimation method. The table shows the summary statistics of the posterior distribution and its 95% HPD credible intervals.

Table 4.11: Posterior summary and credible intervals

Coefficient	Mean	Naive SE	95% HPD credible interval
Intercept	-8.179	0.005	[-8.830 : -7.505]**
Rain	-0.001	<0.001	[-0.003: 0.000]***
Cap(Ref)	- - - -	- - - -	- - - -
Mop	2.169	0.001	[1.994: 2.329] ***
Sekh	0.335	0.002	[0.090 : 0.562]**
Vhem	2.817	0.001	[2.664 : 2.992]***
Watbg	-0.219	0.002	[-0.454 : -0.006]**
2014(Ref)	- - - -	- - - -	- - - -
2015	0.086	0.001	[0.013 : 0.157]**
tn	0.111	<0.001	[0.093 : 0.129]***
td	0.165	<0.001	[0.145: 0.184]**
ele	-0.0004	<0.001	[-0.001: 0.000]*
ndvi	4.911	0.004	[4.343 : 5.457]***
Key:	p <0.001 '***'	0.01 '**'	0.05 '*' ⩾0.05 ' '

All the 95% credible intervals presented in Table 4.11 do not include zero, which indicates that all the parameters are significant. However, the parameter of NDVI is extremely significant, while other parameters are moderately significant. This implies that malaria incidence is affected more by NDVI than other environmental factors.

Both 95% HPD credible intervals for the regression coefficients of the covariates rain and elevation are negative. This implies that there is a very high probability that the estimates of these regression coefficients are negative. Therefore, we can conclude that the relationship between malaria incidence and each of the covariates rain and elevation is negative. That is, an increase in rainfall leads to a decrease in malaria incidence and an increase in elevation above sea level leads to a decrease in malaria incidence.

All the 95% HPD credible intervals for temperature during the night (tn), temperature during the day (td) and NDVI are positive, which indicate that there

is a very high probability that the estimates of these regression coefficients are positive. Therefore, we can conclude that the relationship between each of these covariates and malaria incidence is positive. That is, an increase in temperature during the night, temperature during the day and NDVI results in an increase in malaria incidence.

The 95% HPD credible intervals for Mopani, Sekhukhune and Vhembe districts are positive, which indicates that as malaria incidence increases in each of these districts, it also increases in Capricorn district (Reference variable). However, both the 95% HPD credible intervals for Waterberg are negative. This implies that if malaria incidence increases in Capricorn district, then it decreases in Waterberg district. We can now conclude that according to the MCMC estimation methods applied to obtain the model in Table 4.11, there is a relationship between malaria incidence and each of the environmental factors included in this study.

4.3.3 Comparison of classical (MLE) and Bayesian (MCMC) methods of estimation

The classical framework in this study employed MLE method to estimate the NB model parameters, standard errors, p-values and the 95% confidence intervals. The Bayesian framework employed the MCMC estimation method to estimate the posterior mean, naive standard errors and the 95% HPD credible intervals. The posterior distribution in this case is obtained by developing the NB model through the Bayesian context. We compare the classical estimation method and the Bayesian estimation method using the NB model presented in Table 4.9 and the posterior summary presented in Table 4.11. The parameter estimates of the Bayesian framework in Table 4.11 are all significant, while the parameter estimates of the classical framework in Table 4.9 are significant, except for temperature during the day, elevation, NDVI, Sekhukhune district and

the year 2015. Therefore, to ensure that the comparisons of the MLE method and the MCMC estimation method are unbiased, we have based our comparisons on the parameter estimates that are significant in both the classical and Bayesian frameworks. These are the parameter estimates for rain, temperature during the night, Mopani, Vhembe and Waterberg.

The parameter estimate for rain is moderately significant in both Table 4.9 and Table 4.11. The standard error associated with rain is 0.1012, while its naive standard error is 0.0227. This implies that, based on the covariate rain, the MLE method generates more errors compared to MCMC estimation method. The difference of the upper and the lower limits of the 95% confidence interval of the parameter estimate of rain is -0.1043, while the difference of the upper and the lower limits of its 95% HPD credible interval is -0.0312. Therefore, we can conclude in this case, based on the parameter estimates of rain, that the 95% HPD credible interval is shorter than the 95% confidence interval since its difference is smaller compared to the difference of the upper and lower limits of the confidence interval.

The estimates of the parameters for temperature during the night are significant according to Table 4.9 and Table 4.11. The standard error for temperature during the night is 0.4396, while its naive standard error is 0.0236. Therefore, the standard error is almost 19 times larger than the naive standard error of the same covariate. Hence in this case, the MLE method generates more errors than the MCMC estimation method. The difference of the upper and lower limits of the 95% confidence interval for temperature during the night is 1.0194, while the difference of the upper and lower limits of the 95% HPD credible interval for the same covariate is 0.3820. Therefore, the difference of the limits of the 95% confidence interval is larger than the difference of the limits of the 95% HPD credible interval. Hence we can conclude, in this case, that the cred-

ible interval is narrower than the confidence interval.

The parameter estimates for Mopani are positive in both Table 4.9 and Table 4.11, implying that Mopani district is more susceptible to malaria incidence than Capricorn district, according to both the Bayesian and the classical frameworks. However, the standard error for Mopani is 0.7674, while its naive standard error is 0.0607. This implies that the standard error, in this case, is almost 13 times larger than the naive standard error. The difference of the 95% confidence interval limits for Mopani is 0.8535, while the difference of its 95% HPD credible interval is 0.3345. Hence we can conclude, in this case, that the credible interval is shorter than the confidence interval.

The estimates of the parameters for Vhembe district are positive and significant according to Table 4.9 and Table 4.11, implying that Vhembe district is more susceptable to malaria incidence than Capricorn district, according to both the Bayesian and classical frameworks. The standard error for Vhembe is 0.7773, while its naive standard error is 0.0591. Therefore, the standard error is almost 13 times larger than the naive standard error of the same covariate. Hence in this case, the MLE method generated more errors than the MCMC estimation method. The difference of the upper and lower limits of the 95% confidence interval for Vhembe in Table 4.9 is 0.8728, while the difference of the upper and lower limits of the 95% HPD credible interval for the same covariate in Table 4.11 is 0.3282. Therefore, the difference of the limits of the 95% confidence interval is larger than the difference of the limits of the 95% HPD credible interval. Hence we can conclude, in this case, that the credible interval is narrower than the confidence interval.

According to the classical framework as presented in Table 4.9, the parameter estimate for Waterberg is positive, which indicates that malaria incidence

is more concentrated in Waterberg than in Capricorn district. However, according to the Bayesian framework presented in Table 4.11, the parameter estimate for Waterberg is negative, which indicates that malaria incidence is more concentrated in Capricorn than in Waterberg district. The standard error for Waterberg is 0.8428, while its naive standard error 0.0796. This implies that the standard error in this case is almost 11 times larger than the naive standard error. The difference of the 95% confidence interval limits for Waterberg is 0.9534, while the difference of its 95% HPD credible interval is 0.1557. Hence we can conclude, in this case, that the credible interval is shorter than the confidence interval.

Based on the comparisons of the errors generated by the classical method of estimation presented in Table 4.9 and the errors generated by the Bayesian method of estimation presented in Table 4.11, the classical method of estimation generates more errors than the Bayesian method of estimation. Again, these two tables (Table2 4.9 and 4.11) provide the evidence that the Bayesian estimation method produces credible intervals that are shorter than the confidence intervals produced by the classical method of estimation. Hence, we can conclude that the MCMC estimation method employed in the Bayesian framework, produces better estimations than the MLE method employed in the classical framework.

Chapter 5

Conclusion and recommendations

5.1 Conclusion

In this study, we have modelled malaria incidence in relation to rainfall, temperature, normalised vegetation index (NDVI), elevation and time in quarters from 2014 to 2015 across the various districts of the Limpopo province. Since malaria incidence data are counts, we developed three Poisson models and used the AIC method to select the best model. However, the selected Poisson model was found to be overdispersed. Therefore, to correct for the overdispersed Poisson model we used the negative binomial (NB) model which naturally account for overdispersion. The three Poisson models and NB model developed employed the MLE method for parameter estimation. Hence, these models make up the classical part of this study. The NB model was also developed through the use of MCMC parameter estimation method, which is used to obtain the

posterior distribution. All the inferences for the Bayesian framework are based on the posterior obtained, which make up the Bayesian part of this study.

Both the Bayesian and classical methods revealed a positive relationship between malaria incidence and temperature during the night. That is, an increase in temperature during the night results in an increase in malaria incidence. Therefore, we can conclude that the risk of malaria transmission is high during warm nights, which are usually the nights of summer seasons. The Bayesian and classical frameworks produced similar results about the relationship between malaria incidence and rainfall, which was found to be negative. Therefore, we can conclude that an increase in the amount of rainfall results in a decrease of malaria incidence. The classical framework does not provide any evidence of an existing relationship between malaria incidence and either elevation, temperature during the day nor NDVI. However, the Bayesian framework revealed that an increase in NDVI or temperature during the day lead to increased malaria incidence while an increase in elevation above sea level leads to decreased malaria incidence. The two methods used in this study suggest that if malaria incidence increases in Mopani and/or Vhembe districts, then it also increases in Capricorn district. The classical framework revealed no pattern of malaria incidence between Capricorn and Sekhukhune district while the Bayesian framework suggests that if malaria incidence increases in Sekhukhune district, then it also increases in Capricorn district. The Bayesian framework also suggests that if malaria incidence decreases in Waterberg district then it increases in Capricorn district while in contrast, the classical framework suggest that if malaria incidence increases in Waterberg district then it also increases in Capricorn district. Both methods affirm that Vhembe district is more susceptible to malaria incidence, followed by Mopani district. The classical method did not identify any particular trend of malaria incidence over the period of study. However, the Bayesian method

identified an upward trend of malaria incidence over the period of the study. The MLE method generated more errors and wider intervals while the MCMC estimation method generated fewer errors and narrower intervals. Therefore, we can conclude that the Bayesian method of estimation outperforms the classical method of estimation.

The results of this study are similar to the results of Ramalata (2017), which provided evidence that NB model fit malaria count data better than the Poisson model. The results of this study are also in agreement with the findings of Zayeri et al. (2011) which revealed a negative relationship between rainfall and malaria incidence. Furthermore, the results of this study are similar to the results of Gosoniu et al. (2006) and Shimaponda-Mataa et al. (2017), which revealed a positive relationship between malaria risk and temperature during the night. Gosoniu et al. (2006) also obtained similar results as this study, which identified a positive relationship between malaria incidence and NDVI. In agreement with the results of this study, Gerritsen et al. (2008) identified Vhembe to be the district that is more susceptible to malaria incidence compared to other districts of Limpopo province.

5.2 Recommendations

We recommend that the Department of Health and Malaria Control Programme of South Africa allocate more resources for malaria prevention, control and elimination to Vhembe and Mopani districts of Limpopo province. We also recommend that the government provide educational seminars to educate the South African communities on how to prevent malaria transmission, especially during the warm summer nights.

5.3 Future research

Prior distribution of a parameter is the probability that represents one's uncertainty about the parameter before the data are examined (Institute, 2014). The product of the prior distribution and the maximum likelihood function gives us the posterior distribution, which is used to carry out all the Bayesian inferences. Therefore, the strength of a posterior distribution depends on the strength of the prior distribution and the magnitude of the data available. According to Kass and Wasserman (1996), most Bayesian analyses are performed using the noninformative priors constructed by some formal rule. Franck et al. (2019) outlined that researchers usually choose prior classes based on the goals of their study. However, selecting the prior distributions without considering the prior selection methods may result in the construction of an improper posterior distribution which cannot be used for inferences. This study only discussed the noniformative and informative priors. However, future research may involve studies on the methods to select the best prior distributions and also examine the improper, conjugate and Jeffrey's priors in detail.

References

BLUMBERG, L. AND FREAN, J. (2017). Malaria reduces globally but rebounds across Southern Africa. *Southern African Journal of Infectious Diseases*, **32** (2), 3–4.

BOATENG, A. (2012). Modeling the occurrence and incidence of malaria cases: a case study at Obuasi Government Hospital. An Aphil thesis submitted to Kwame Nkrumah University of Science and Technology (KNUST), Ghana.

BOLSTAD, W. M. (2010). *Understanding computational Bayesian statistics*, volume 644. John Wiley & Sons. Hamilton, New Zealand.

CONSUL, P. C. AND JAIN, G. C. (1973). A generalization of the Poisson distribution. *Technometrics*, **15** (4), 791–799.

COX, S. N., GUIDERA, K. E., SIMON, M. J., NONYANE, B. A. S., BRIEGER, W., BORNMAN, M. S., AND KRUGER, P. S. (2018). Interactive malaria education intervention and its effect on community participant knowledge: The Malaria Awareness Program in Vhembe District, Limpopo, South Africa. *International Quarterly of Community Health Education*, **38** (2), 147–158.

DUVALL, R. (1999). A Bayesian approach to negative binomial parameter estimation. *In Casualty Actuarial Society Forum*. pp. 377–85.

FRANCK, C. T., KOFFARNUS, M. N., MCKERCHAR, T. L., AND BICKEL, W. K. (2019). An overview of bayesian reasoning in the analysis of delay-discounting data. *Journal of the experimental analysis of behavior*.

GERRITSEN, A. A., KRUGER, P., VAN DER LOEFF, M. F. S., AND GROBUSCH, M. P. (2008). Malaria incidence in Limpopo Province, South Africa, 1998–2007. *Malaria Journal*, **7** (1), 162.

GOSONIU, L., VOUNATSOU, P., SOGOBA, N., AND SMITH, T. (2006). Bayesian modelling of geostatistical malaria risk data. *Geospatial Health*, **1** (1), 127–139.

HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications, Oxford University Press, New York.

HAY, S., SNOW, R., AND ROGERS, D. (1998). From predicting mosquito habitat to malaria seasons using remotely sensed data: practice, problems and perspectives. *Parasitology Today*, **14** (8), 306–313.

HILBE, J. M. (2011). *Negative binomial regression*. Cambridge University Press, New York.

INSTITUTE, S. (2014). *SAS 9.4 Output delivery system: User's guide*. SAS Institute, South Africa.

KASS, R. E. AND WASSERMAN, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, **91** (435), 1343–1370.

KAZEMBE, L. N. (2007). Spatial modelling and risk factors of malaria incidence in northern Malawi. *Acta Tropica*, **102** (2), 126–137.

KLEINSCHMIDT, I., SHARP, B., CLARKE, G., CURTIS, B., AND FRASER, C. (2001). Use of generalized linear mixed models in the spatial analysis of small-area malaria incidence rates in Kwazulu-Natal, South Africa. *American Journal of Epidemiology*, **153** (12), 1213–1221.

LI, H. (2008). *Bayesian hierarchical models for spatial count data with application to fire frequency in British Columbia, University of Victoria*. Ph.D. thesis.

LINDSEY, H. L. (2011). An introduction to Bayesian Methodology via WinBUGS and PROC MCMC, Brigham Young University-Provo.

LIO, Y. (2009). A note on Bayesian Estimation for the Negative-Binomial Model. *Pliska Studia Mathematica Bulgarica*, **19** (1), 207p–216p.

MAZEROLLE, M. (2006). Improving data analysis in herpetology: Using Akaike's Information Criterion (AIC) to assess the strength of biological hypotheses. *Amphibia-Reptilia*, **27** (2), 169–180.

MOIROUX, N., BOUSSARI, O., DJÈNONTIN, A., DAMIEN, G., COTTRELL, G., HENRY, M.-C., GUIS, H., AND CORBEL, V. (2012). Dry season determinants of malaria disease and net use in Benin, West Africa. *PLoS One*, **7** (1), e30558.

NAJERA, J. A., KOUZNETSOV, R., DELACOLLETTE, C., AND ORGANIZATION, W. H. (1998). Malaria epidemics: detection and control, forecasting and prevention. Technical report, Geneva: World Health Organization.

OMONIJO, A., MATZARAKIS, A., OGUNTOKE, O., AND ADEOFUN, C. (2011). Influence of weather and climate on malaria occurrence based on human-biometeorological methods in Ondo State, Nigeria. *Journal of Environmental Science and Engineering*, **5** (9).

PLONSKY, L. AND OSWALD, F. L. (2017). Multiple regression as a flexible alternative to ANOVA in L2 research. *Studies in Second Language Acquisition*, **39** (3), 579–592.

RACHEV, S. T., HSU, J. S., BAGASHEVA, B. S., AND FABOZZI, F. J. (2008). *Bayesian methods in finance*, volume 153. John Wiley & Sons.

RAMALATA, A. (2017). Analysis of malaria risk factors in the Limpopo Province, South Africa: An application of poisson and negative binomial regression models, MSc dissertation, University of Limpopo, South Africa.

RAMAN, J., MORRIS, N., FREAN, J., BROOKE, B., BLUMBERG, L., KRUGER, P., MABUSA, A., RASWISWI, E., SHANDUKANI, B., AND MISANI, E. (2016). Reviewing South Africa's malaria elimination strategy (2012–2018): progress, challenges and priorities. *Malaria Journal*, **15** (1), 438.

REDDY, L. H., ARIAS, J. L., NICOLAS, J., AND COUVREUR, P. (2012). Magnetic nanoparticles: design and characterization, toxicity and biocompatibility, pharmaceutical and biomedical applications. *Chemical reviews*, **112** (11), 5818–5878.

SACHS, J. AND MALANEY, P. (2002). The economic and social burden of malaria. *Nature*, **415** (6872), 680.

SCHMIDT, S. (2017). Travel to malaria areas. What's the fuss about the buzz? *SA Pharmacist's Assistant*, **17** (4), 27–30.

SCOTT, J. AND PILLOW, J. W. (2012). Fully Bayesian inference for neural models with negative-binomial spiking. *In Advances in Neural Information Processing Systems*. pp. 1898–1906.

SHIMAPONDA-MATAA, N. M., TEMBO-MWASE, E., GEBRESLASIE, M., ACHIA, T. N., AND MUKARATIRWA, S. (2017). Modelling the influence of temperature and rainfall on malaria incidence in four endemic provinces of Zambia using semiparametric Poisson regression. *Acta Tropica*, **166**, 81–91.

SNOW, R. W. (2015). Global malaria eradication and the importance of plasmodium falciparum epidemiology in Africa. *BMC Medicine*, **13** (1), 23.

SPOTTISWOODE, N., DUFFY, P. E., AND DRAKESMITH, H. (2014). Iron, anemia and hepcidin in malaria. *Frontiers in Pharmacology*, **5**, P.125.

TANG, W., HE, H., AND TU, X. M. (2012). *Applied categorical and count data analysis*. CRC Press, Boca Raton New York.

TURNER, H. (2008). Introduction to generalized linear models. *Rapport Technique, Vienna University of Economics and Business*.

YÉ, Y., LOUIS, V. R., SIMBORO, S., AND SAUERBORN, R. (2007). Effect of meteorological factors on clinical malaria risk among children: an assessment using village-based meteorological stations and community-based parasitological survey. *BMC Public Health*, **7** (1), 101.

ZAYERI, F., SALEHI, M., AND PIRHOSSEINI, H. (2011). Geographical mapping and Bayesian spatial modeling of malaria incidence in Sistan and Baluchistan province, Iran. *Asian Pacific Journal of Tropical Medicine*, **4** (12), 985–992.

ZYPHUR, M. J. AND OSWALD, F. L. (2015). Bayesian estimation and inference: A user's guide. *Journal of Management*, **41** (2), 390–420.

Appendix

SOME SELECTED R CODES

R codes for descriptive statistics

```
rm(list = ls())

library(car)
library(multcomp)
library(lmtest)

tt=read.table(file.choose(),header=T)
tt
head(tt,4)
attach(tt)
summary(tt)

datause=data.frame(mal,dist,pop,ele,tn,td,ndvi,rain,dyear)
attach(datause)
box plots
par(mfrow=c(2,2))
plot(as.factor(dist),mal,xlab="Districts",ylab="Malaria counts", main="Malaria
counts versus districts",cex.main=1.2)
plot(as.factor(dyear),mal,xlab="Years",ylab="Malaria counts",main="Malaria counts
```

```
versus years",cex.main=1.2)
plot(rain,mal,xlab="Rainfall",ylab="Malaria counts",main="Malaria counts ver-
sus rainfall",cex.main=1.2)
plot(tn, mal,xlab="Temperature during the night",ylab="Malaria
counts",main="Malaria counts versus temperature during the night",cex.main=1.2)
plot(td, mal,xlab="Temperature during the day",ylab="Malaria counts",main="Malaria
counts versus temperature during the day",cex.main=1.2)
plot(ele,mal)
plot(ndvi,mal,xlab="Normalised difference vegetation index",ylab="Malaria
counts",main="Malaria counts versus normalised difference vegetation index",cex.main=1.2)
```

R codes for classical models

The Poisson model with all the covariates

```
pm1=glm(mal rain+as.factor(dist)+as.factor(dyear)+tn+td+ele+ndvi+
offset(log(pop)),family="poisson",data=datause)
summary(pm1)
plot(pm1)
coef(pm1)
confint(pm1)
exp(cbind(C0=coef(pm1),confint(pm1)))
anova(pm1)
residuals(pm1)
predict(pm1)
```

The Poisson model in exclusion of the district explanatory variable

```
pm2=glm(mal rain+as.factor(dyear)+td+tn+ele+ndvi+offset(log(pop)),
family="poisson",data=datause)
```

```
summary(pm2)
plot(pm2)
coef(pm2)
confint(pm2) exp(cbind(C0=coef(pm2),confint(pm2))) anova(pm2)
residuals(pm2)
predict(pm2)
```

The Poisson model in exclusion of the NDVI explanatory variable

```
pm3=glm(mal rain+as.factor(dist)+as.factor(dyear)+td+tn+ele
+offset(log(pop)),family="poisson",data=datause)
summary(pm3)
plot(pm3)
coef(pm3)
confint(pm3)
exp(cbind(C0=coef(pm3),confint(pm3)))
anova(pm3)
residuals(pm3)
predict(pm3)
```

Detection of overdispersion

```
require(stats)
mean(mal)
var(mal)

install.packages("qcc")
install.packages("car")
install.packages("lmtest")
```

```
install.packages("multcomp")
install.packages("AER")

library(qcc)
qcc.overdispersion.test(tt$mal, type="poisson")
qcc.overdispersion.test(mal, type="poisson")
```

NB model

```
require(MASS)

pm5=glm.nb(mal tn+as.factor(dist)+ele+
offset(log(pop))+rain+as.factor(dyear))
summary(pm5)
par(mfrow=c(2,2))
plot(pm5)
coef(pm5)
confint(pm5)
exp(cbind(C0=coef(pm5),confint(pm5)))
anova(pm5)
residuals(pm5)
predict(pm5)
exp(cbind(C0=coef(pm5),confint(pm5)))
```

Bayesian methods

Data definition

```
data1=tt
```

```
head(data1)
attach(data1)
data1=data.frame(mal,dist,pop,ele,tn,td,ndvi,rain,dyear)
```

The development of NB using MCMC

```
require(MCMCpack)
posterior <- MCMCnegbin(mal rain+as.factor(dist)+tn+td+as.factor(dyear)+ele+
ndvi+offset(log(pop)), b0=0, B0 = 0.1, sigma.mu = 5, sigma.var = 25, data=data1,
verbose=1000, burnin = 5000, mcmc=10000, thin=2)
```

Posterior summary and the convergence of the Markov chain

```
summary(posterior)
par(mar = rep(2, 4))
plot(posterior)
```