Modeling the Impact of Climatic Variables on Malaria Incidences: A Case Study of Apac District, Northern Uganda

AYO EUNICE

MASTER OF SCIENCE

Mathematics-Statistics Option

Pan African University Institute for Basic Sciences, Technology and Innovation

2018

Modeling the Impact of Climatic Variables on Malaria Incidences: A Case Study of Apac District, Northern Uganda

AYO EUNICE

MS 300-0001/16

A Research Thesis submitted to Pan African University Institute of Science, Technology and Innovation in partial fulfillment of the requirement for the award of the degree of Master of Science in Mathematics (Statistics Option) of the Pan African University

2018

Declaration

I hereby declare that this thesis is my original work and has not been presented for a degree in any other University.

AYO EUNICE MS 300-0001/16

Signature..... Date.....

Supervisors declaration

I hereby declare that the preparation and presentation of this thesis was supervised in accordance with the guidelines on supervision of thesis laid down by the University.

1. Dr. Wanjoya Anthony Department of Statistics and Actuarial Sciences Jomo Kenyatta University of Agriculture and Technology P.O Box 62000-002000, Nairobi, Kenya

   Signature..... Date.....

2. Prof. Livingstone Luboobi Institute of Mathematical Sciences Strathmore University P.O Box 59857-00200, Nairobi, Kenya

   Signature..... Date.....

Acknowledgment

I would like to give thanks to God who has enabled me reach this far. I would like to thank my supervisors Dr Anthony Wanjoya and Prof Livingstone Luboobi for their generous support, encouragement, and many invaluable suggestions and comments for this research. I would like to thank PAUSTI for giving me this opportunity to climb up the academic ladder. I wish to thank all my family, especially my husband Mr Odongo for his love, patience and sacrifice in the pursuit of my goals. I wish to also thank my parents, my daughter, Gabriella Precious for their love and affection. Finally, I wish to thank all my classmates and friends who have supported me during my coursework and my stay in Kenya.

Dedication

I humbly dedicate this piece of work to my father, Mr Okello Celestine.

Abstract

Malaria is a major cause of morbidity and mortality in Apac District, Northern Uganda. Hence, the study aimed to model malaria incidences with respect to climate variables for the period 2007 to 2016 in Apac District. Data on monthly Malaria incidence in Apac District for the period January 2007 to December 2016 was obtained from the Ministry of Health, Uganda whereas climate data was obtained from Uganda National Meteorological Authority. Generalized linear models, Poisson and negative binomial regression models were employed to analyze the data. These models were used to fit monthly malaria incidences as a function of monthly rainfall and average temperature. Negative binomial model provided a better fit as compared to the Poisson regression model as indicated by the residual plots and residual deviances. The Pearson correlation test indicated a strong positive association between rainfall and Malaria incidences. The Autoregressive integrated moving average, ARIMA $(1,0,0)(1,1,0)^{12}$ was found to be the best fit model for the malaria time series data. ARIMA models for time series analysis was found to be a simple and reliable tool for producing relaible forecasts for malaria incidences in Apac District, Uganda. This study showed a significant association between monthly malaria incidence and climate variables that is rainfall and temperature. This study provided useful information for predicting malaria incidence and developing the future warning system. This is an important tool for policy makers to put in place effective control measures for malaria early enough. Malaria still remains a public health concern in Uganda, in particular Apac District.

List of Acronyms

AIC : Akaike Information Criterion

AICc : Akaike Information Criterion corrected

ACF : Autocorrelation Function

BIC : Bayesian Information Criterion

ADF : Augmented Dickey-Fuller

EIR : Entomological Inoculation Rate

VIF : Variance Inflation Factor

AR : Autoregressive

EF : Exponential Family

MA : Moving Average

ARIMA : Autoregressive Integrated Moving Average

SARIMA : Seasonal Autoregressive Integrated Moving Average

ARCH : Autoregressive Conditional Heteroscedasticity

df : Degrees of freedom

PACF : Partial Autocorrelation Function

MAE : Mean Absolute Error

MLE : Maximum Likelihood Estimation

OLS : Ordinary Least Squares

GLM : Generalized Linear Models

Contents

List of Figures

List of Tables

1. Introduction

1.1. Background of Study

Uganda is one of the Sub-Saharan African countries where malaria is still endemic in over 90% of the country's regions (Health, 2014; Talisuna et al., 2007; Yeka et al., 2012). According to the National Malaria Control Programme, malaria alone has shown to contribute to between 30 and 50% of outpatient visits, $15 - 20\%$ of hospital admissions and 20% of hospital deaths with most of this burden found in children under 5 years and pregnant women (Health, 2014). Transmission of malaria is very complicated. The impact of climatic variables on malaria patterns still remains controversial. The aim of this study is to model malaria incidences in Apac District, Northern Uganda with respect to climate variables specifically rainfall and temperature.

The greatest burden of malaria, remains in the heart land of Africa, characterized by limited infrastructure to monitor disease trends, large contiguous areas of high transmission and low coverage of control interventions. The epidemiology of malaria varies widely in Uganda, from highland regions with low prevalence and unstable disease to large regions with dense agricultural settlement and some of the highest recorded malaria intensities in the world (Okello et al., 2006).

The climate in Uganda allows stable, year round malaria transmission with relatively little seasonal variability in most areas. Malaria is highly endemic in Uganda with some of the highest recorded Entomological Innoculation Rates (EIR, infective mosquito bites per person per year) in the world, including rates of 1586 in Apac district and 562 in Tororo District measured in 2001 to 2002 (Okello et al., 2006).

Malaria remains one of the leading health problems of the developing world, and

Uganda bears a particularly large burden from the disease. Our understanding is limited by a lack of reliable data, but it is clear that the prevalence of malaria infection, incidence of disease, and mortality from severe malaria all remain very high. Uganda has made progress in implementing key malaria control measures, in particular distribution of insecticide impregnated bednets, indoor residual spraying of insecticides, utilization of artemisinin-based combination therapy to treat uncomplicated malaria, and provision of intermittent preventive therapy for pregnant women. However, despite enthusiasm regarding the potential for the elimination of malaria in other areas, there is no convincing evidence that the burden of malaria has decreased in Uganda in recent years. Major challenges to malaria control in Uganda include very high malaria transmission intensity, inadequate health care resources, a weak health system, inadequate understanding of malaria epidemiology and the impact of control interventions, increasing resistance of parasites to drugs and of mosquitoes to insecticides, inappropriate case management, inadequate utilization of drugs to prevent malaria, and inadequate epidemic preparedness and response. Despite these challenges, prospects for the control of malaria have improved, and with attention to underlying challenges, progress toward the control of malaria in Uganda can be expected (Krefis et al., 2011).

The relationship between climatic variables and malaria transmission has been reported in many countries (Aribodor et al., 2016). A recent resurgence of malaria in the East African highlands involves multiple factors; climate and land use change, drug resistance, variable disease control efforts, and other socio-demographic factors (Pascual et al., 2006). Malaria is an extremely climate-sensitive disease (Rogers & Randolph, 2000) common in the tropics ,but also reported in mild-to-cold climates (Hulden & Hulden, 2009).

Rainfall and temperature anomalies are widely considered to be a major driver of inter-annual variability of malaria incidence in the semi-arid areas of Africa (Connor et al., 1999), and recently recorded a warming trend in the East African Highlands that corresponded with concomitant increases in malaria incidences (Pascual et al., 2006).

Based on the background study of malaria above, the impact of weather and environmental factors on dynamics of malaria has attracted considerable attention in recent

years, yet uncertainties around future disease trends under environment change remain. The role of climate as a driving force for malaria incidences is still a subject of considerable attention (Shanks et al., 2002). Assessing the impact of climate variables on malaria incidences is challenging because of a high spatial climate variability and lack of a long term data series on malaria cases from different hospitals. Temperature affects the development rates and survival of malaria parasites and mosquito vectors. Rainfall influences the availability of the mosquito larvae habitats and hence a breeding ground for mosquitoes. Temperature and rainfall may have synergistic effects on the transmission of malaria. Therefore, there is need to analyze the simultaneous effects of rainfall and temperature on malaria incidences. However, the association between climate variables and malaria incidences in Apac District has not been studied.

Malaria has historically been a very serious health problem and currently poses the most significant threat to the health of the people in malaria prone areas. Uganda show that more than 55 percent of pediatric cases are due to malaria (Martens & Hall, 2000). Malaria currently accounts for: 25% of all out patients visits at health facilities,20% of hospital admissions, 9.14% of in patient deaths,a case fatality rate of 35% (which is an under estimate, since many malaria cases go unreported especially those in areas in accessible to health facilities), 23% and 11% of deaths among the under fives in high and medium malaria transmission areas respectively (estimates by Uganda Ministry of Health), severe malarial anaemia is responsible for a case fatality rate of 8.25% among pediatric admissions.

Annually, more than one billion people are infected and more than one million die from vector borne diseases. World Health Organization (WHO) has highlighted the serious and increasing threat of vector-borne diseases with the theme 'preventing vector-borne diseases' and also with the slogan "small bite, big threat" for the year 2014. Among vector-borne diseases, malaria poses the biggest threat with about 40% of the world's population at risk of infection. In 2013, 97 countries had ongoing transmission of malaria (WHO, 2014).

Malaria decreases economic growth by more than one percentage point per year in endemic countries. Malaria transmission usually coincides with the harvesting

season and brief periods of illness exact a high cost on the world's poorest regions (Millennium, 2005).

Malaria still remains a public health problem in developing countries and changing environmental and climatic conditions are considered as the biggest challenge in fighting against the scourge of malaria (McMichael et al., 2006). Malaria is an entirely preventable and treatable illness caused by parasites of plasmodium species and transmitted by the bites of Anopheles mosquito. Although preventable and treatable, malaria causes significant morbidity and mortality, particularly in poor regions. In 2015, 91 countries and areas had ongoing malaria transmission. Malaria is preventable and curable, and increased efforts are dramatically reducing the malaria burden in many places. Between 2010 and 2015, malaria incidence among populations at risk (the rate of new cases) fell by 21% globally. In that same period, malaria mortality rates among populations at risk fell by 29% globally among all age groups, and by 35% among children under 5. The WHO African Region carries a disproportionately high share of the global malaria burden. In 2015, the region was home to 90% of malaria cases and 92% of malaria deaths (WHO, 2017) .

Due to severe health impact of malaria, there is a growing need for methods that will allow forecasting and early warning with timely detection in areas of unstable transmission, so that more control measures can be implemented effectively (WHO, 2004). Studies of malaria epidemics have shown their association with excess rainfall, temperature and vegetation density. This is observed in the direct correlation between an abundance of Anopheles mosquitoes and rainfall, increased transmission and temperature (Gomez-Elipe et al., 2007).

1.2. Statement of the problem

Transmission of malaria is very complicated.The impact of climatic variables on malaria patterns still remains controversial. Therefore, there is need for a useful model that is able to assess the impact of climatic variables on malaria incidences and also forecast the malaria incidences. This is helpful in the development of reliable malaria warning

systems.

The number of malaria cases in relation to meteorological factors that is rainfall and temperature is still unknown. It is also probable that the efforts to reduce malaria do not specifically take into account the meteorological factors likely to aggravate malaria disease.

Applying linear regression to count data leads to inconsistent standard errors and may produce negative predictions for the dependent variable. Even with a logged dependent variable, the least squared estimates have these problems and are biased and inconsistent. Therefore count dependent variables require different modeling. The most common assumption of count data distribution is the Poisson distribution which restricts the data distribution to be equal-dispersion (the conditional variance equals the conditional mean). This stringent restriction cannot handle many empirical applications. Other modeling distributions have been developed. Negative binomial distributions have been widely used in situations where counts display over-dispersion (conditional variance exceeds the conditional mean).

1.3. Research objectives

Based on the statement of the problem above, the objectives of the study are as follows:

General objective

The main purpose of the study is to model the impact of climate variables on malaria incidences in Apac District, Northern Uganda.

Specific objectives

These are:

1. To analyze the effects of rainfall on expected malaria incidences.

2. To analyze the effects of average temperature and on expected malaria incidences.

3. To analyze the effects of rainfall and average temperature and on expected malaria incidences.

4. To forecast malaria incidences using an appropriate statistical method.

1.4. Significance of the study

Malaria is a disease that is constantly changing. Malaria affects the health and wealth of nations and individuals. In Africa today, malaria is understood to be both a disease of poverty and a cause of poverty (Gomez-Elipe et al., 2007). Malaria has significant measurable direct and indirect costs, and has been shown to be a major constraint to economic development (Sachs & Malaney, 2002). This means the gap in prosperity between countries with malaria and countries without malaria has become wider every single year.

Forecasting future malaria incidences is important for policy makers in public health planning in development and enhancement of early warning systems for malaria in the different regions. Knowing the annual trend of malaria given the annual trend of rainfall is important to public health professionals to assist in health care assessments, service planning and policy development in regards to malaria incidences in the different regions.

(Sachs & Malaney, 2002) showed that where malaria has been eliminated, economic growth has increased substantially. Hence the importance to develop a statistical model that enables us assess the impact of climatic variables on malaria patterns which enables policy makers to develop targeting, preventative and control strategies which is cost effective. The Roll Back Malaria (RBM) Initiative was launched in 1998 with the aim to markedly reduce malaria morbidity and mortality. In the year 2000, the world launched Millennium Development Goals (MDGs) and Goal 6C was to halt and reverse the incidence of malaria by 2015.

Some methods in literature reviewed give contradicting results on the malaria patterns hence the need to develop better models that can give good inferences on the malaria patterns. Following the end of MDG, the World Health Organization member states, Uganda inclusive, on 20th May, 2015 agreed on a new global malaria strategy for $2016 - 2030$. The strategy aims to reduce the global disease burden by 40% by 2020, and by at least 90% by 2030. It also aims to eliminate malaria in at least 35 new countries by 2030.

2. Literature Review

2.1. Malaria models

Several studies have been carried out on malaria incidence. Nkurunziza et al. (2010), investigated the effects of climate on malaria in Burundi using generalized linear models and generalized additive mixed models. The results suggest a strong positive association between malaria incidence in a given month and minimum temperature of the previous month. In contrast, it was found that rainfall and maximum temperature in a given month have possible negative effect on malaria incidence of the same month.

Lindsay et al. (2000), compared the level of malaria infection in children from 22 communities in an area of unstable transmission in the Usambara mountains, Tanzania, immediately before and after one of the strongest recorded El Niño Southern Oscillation events. They found strikingly less malaria than in the preceding year despite 2.4 times more rainfall than normal resulted from the event.

Nkurunziza et al. (2011), used semi-parametric regression models to model the dependence of malaria cases on spatial determinants and climatic covariates including rainfall, temperature and humidity in Burundi. The results obtained suggested that malaria incidence in a given month is strongly associated with minimum temperature of the previous months.

Huang et al. (2011), modeled separate meteorological factors, the model with rainfall performed better than the models with other factors respectively. The results showed that the way rainfall influenced malaria incidence was different from other factors, which could be interpreted as rainfall having a greater influence than other factors.

Zhou et al. (2004), used non linear mixed-regression model to investigate the association between auto regression (number of malaria inpatients during previous time period), seasonality and climate variability, and the number of monthly malaria inpatients of the past 10 to 20 years in seven highland sites in East Africa. The model did not take into consideration other important factors that also impact on malaria incidences for example; topography, human settlement pattern, land use, and drug resistance.

Patience and Osagie (2014), studied the trend of malaria prevalence in Minna, Nigeria, by employing Poisson and Negative binomial regression models. The results revealed that the prevalence of malaria is still on increase by 6% on monthly basis. Musa et al. (2012), examined the relationship between malaria and environmental and socioeconomic variables in the Sudan using health production modified model. The regression results showed significant relationships between malaria and rainfall and water bodies. Other variables including Human Development Index, temperature, population density and percent of cultivated areas were not significant.

Kakchapati and Ardkaew (2011), carried out a study to identify the spatial and trends of malaria incidence in Nepal Poisson and negative binomial regression models were used to fit malaria incidence rates as a function of year and location. The study showed a steady decreasing trend in malaria incidence, but the numbers of cases are still very high.

Wardrop et al. (2013), studied malaria incidence over time and its association with temperature and rainfall in four counties of Yunnan province, China. Seasonal trend decomposition was used to examine secular trends and seasonal patterns in malaria incidence, a Poisson regression with Distributed lag non-linear models were used to estimate the weather drivers of malaria seasonality. The study revealed that there was a declining trend in malaria incidence in all four counties. Chanda et al. (2013), conducted a study on malaria vector control in South Sudan. The study revealed that the peak of malaria transmission season lasting 7 to 8 months of the year south of the country and 5 to 6 months in the north.

Dræbel et al. (2013), carried out a study using logistic regression to estimate and assess malaria prevalence and the use of malaria risk reduction measures and their as-

sociation with selected background characteristics in South Sudan. The results suggest that educational attainment need not be very advanced to affect practices of malaria prevention and treatment. Primary school attendance was a stronger predictor for use of malaria risk reduction measures than any other selected background characteristics.

Nath and Mwchahary (2013), analyzed the temporal correlation between malaria incidence and climatic variables using malaria incidence rates in Kokrajhar district of Assam over the period 2001 to 2010. Linear regressions were used to obtain linear relationships between climatic factors and malaria incidence. Temperature was found to be negatively correlated with non-forest malaria incidence while relative humidity was positively correlated with forest malaria incidence.

Teklehaimanot et al. (2004), found that malaria was associated with rainfall and minimum temperature in Ethiopia. Daily average number of cases was modeled using a robust Poisson regression with rainfall, minimum temperature and maximum temperatures as explanatory variables in a polynomial distributed lag model in 10 districts of Ethiopia. To improve reliability and generalizability within similar climatic conditions, the districts were grouped into two climatic zones, hot and cold. In cold districts, rainfall was associated with a delayed increase in malaria cases, while the association in the hot districts occurred at relatively shorter lags. In cold districts, minimum temperature was associated with malaria cases with a delayed effect. In hot districts, the effect of minimum temperature was non-significant at most lags, and much of its contribution was relatively immediate.

Sriwattanapongse et al. (2011), used Spearman's correlation between weekly climatic variables (temperatures, relative humidity and rainfall) and malaria to analyze the bi variate relationships between types of malaria parasites and potential climatic factors. A discrete poisson model was used to identify purely spatial clusters of malaria incidence in the high risk areas. A poisson regression model combined with distributed lag non-linear model was used to examine the effects of temperature, relative humidity and rainfall on the number of malaria cases. The residuals were checked to evaluate the adequacy of the model. Sensitivity analysis was performed to ensure that the associations between climate variables and malaria incidences did not change substantially

when the degrees of freedom for climate variables were changed.

Kim et al. (2012), estimated the effects of climate factors on p.vivax malaria transmission using Generalized linear Poisson models and distributed lag non linear models. Their findings suggested that malaria transmission in temperate areas was highly dependent on climate factors.

Zhou et al. (2004), used negative binomial regression model to examine how spatial distribution of the disease changes with inter annual variability of temperature. To analyze the variation in incidence with temperature and altitude, a negative binomial regression to the monthly cases was fitted. Covariates included season, altitude and linearly de-trended temperature (lagged by 3 months).

Muggeo (2008), presented a model for estimation of temperature effects on mortality that is able to capture jointly the typical features of every temperaturedeath relationship, that is, nonlinearity and delayed effect of cold and heat over a few days. Using a segmented approximation along with a doubly penalized spline-based distributed lag parameterization, estimates and relevant standard errors of the cold- and heat-related risks and the heat tolerance are provided. The model is applied to data from Milano, Italy.

Krefis et al. (2011), investigated temporal associations between weekly malaria incidence in 1,993 children $<$15 years of age and weekly rainfall. A time series analysis was conducted by using cross-correlation function and autoregressive modeling. The regression model showed that the level of rainfall predicted the malaria incidence after a time lag of 9 weeks (mean = 60 days) and after a time lag between one and two weeks. The analyses provide evidence that high-resolution precipitation data can directly predict malaria incidence in a highly endemic area. Such models might enable the development of early warning systems and support intervention measures.

2.2. Forecasting Models

Paul et al. (2013), used ARIMA models to forecast average daily price indices of the data series of square Pharmaceuticals Limited (SPL). They found ARIMA$(2, 1, 2)$ as

the best model for forecasting the SPL data series.

Gomez-Elipe et al. (2007), developed an ARIMA model to forecast malaria incidence based on monthly case reports and environmental factors in Karuzi, Burundi. V. Kumar et al. (2014), used ARIMA model to forecast malaria cases using climatic factors as predictors in Delhi, India.
ARIMA$(0,1,1)(0,1,0)^{12}$ was found to be the best fit model.

Wangdi et al. (2010), found ARIMA$(2,1,1)(0,1,1)^{12}$ to be the best possible model to predict malaria cases in Bhutan. The method of ARIMAX modelling was employed to determine predictors of malaria of the subsequent month. ARIMA model was also used for forecasting malaria cases in Sri Lanka (Briet et al., 2008) and Ethopia (Abeku et al., 2002).

Tsitsika et al. (2007), used ARIMA model to forecast pelagic fish production.
They found ARIMA$(1,0,1)$ and ARIMA$(0,1,1)$ to be the best fit models to forecast pelagic fish production. K. Kumar et al. (2004), used ARIMA model to forecast daily maximum surface Ozone concentrations in Brunei Darussalam. They found ARIMA$(1,0,1)$ to be the most suitable model for the surface Ozone data collected at the airport in Brunei Darussalam.

Contreras et al. (2003), used ARIMA models to predict next day electricity prices. They found two ARIMA models to predict hourly prices in the electricity markets of Spain and California. The Spanish model required five hours to predict future electricity prices whereas the Californian model required only two hours for future prediction of electricity prices. Al-Zeaud (2011), used ARIMA model in forecasting volatility. He found ARIMA$(2,0,2)$ to be the best fit model.

Uko and Nkoro (2012), examined the relative predictive power of ARIMA, ECM and VAR models in forecasting inflation in Nigeria. The result showed ARIMA model to be a good predictor of inflation in Nigeria and served as a benchmark model in inflation modeling. Datta (2011), used ARIMA model in forecasting inflation in Bangladesh Economy. He showed that ARIMA$(1,0,1)$ model was the best fit for Bangladesh inflation data.

Liu et al. (2011), used ARIMA model in forecasting incidence of hemorrhagic fever

with renal syndrome in China. ARIMA $(0, 3, 1)$ was found to be the best fit model. From the literature reviewed, in this study Negative binomial regression is considered to analyze the significance of climate variables on malaria incidences. ARIMA models have been used to forecast malaria incidences.

3. Methodology

3.1. Background Statistical Methods

Count data regression models can be represented and understood using generalized linear models (GLM) framework (McCullagh & Nelder, 1992). Poisson regression is commonly used for modeling the number of cases of disease in a specific population within a certain time interval. The Poisson regression is a member of a class of generalized linear models, which is an extension of traditional linear models that allows the mean of a population to depend on a linear predictor through a non linear link function and allows the response probability distribution to be any member of the exponential family distributions. Poisson regression is a special case of (GLM) where the response variable follows Poisson distribution. Poisson models for disease counts are often over-dispersed hence the need for a model which appropriately handles over dispersion in which case negative binomial is considered (Venables & Ripley, 2002). The Negative binomial model is an extension of Poisson model for incidence rates that allows for the over dispersion that commonly occurs for disease count. The Poisson probability distribution is specifically suited for count data, with density function;

$$f(Y_i) = \frac{\lambda^y e^{-\lambda}}{y!}, y = 0, 1, 2, ..., \lambda > 0 \tag{3.1}$$

$f(Y)$ is the probability that the discrete random variable Y takes non- negative integer values, λ is the parameter of the Poisson distribution. It can be proved that;

$$E(Y) = Var(Y) = \lambda \tag{3.2}$$

as in the appendix. A unique feature of Poisson distribution is that the mean is equal to the variance. This is called the equidispersion property of the Poisson distribution. In the Poisson regression model, the number of events y has a Poisson distribution with conditional mean that depends on an individual's characteristics:

$$\lambda_i = E(y_i/x_i) = exp(x_i\beta) \tag{3.3}$$

$$Log(\lambda_i) = x_i\beta \tag{3.4}$$

This is the model for analyzing count data. Under this model as λ_i increases, the conditional variance of y increases. The Poisson regression model can be thought of as a non-linear model (Williams, 2016).

The Negative binomial regression model allows the conditional variance of y to exceed the conditional mean. The mean λ is replaced with the random variable $\tilde{\lambda}$:

$$\tilde{\lambda_i} = exp(x_i\beta + \varepsilon_i) \tag{3.5}$$

where ε is a random error that is assumed to be uncorrelated with x.

$$\tilde{\lambda_i} = exp(x_i\beta)exp(\varepsilon_i) = \lambda_i exp(\varepsilon_i) = \lambda_i\delta_i \tag{3.6}$$

The assumption is that δ has a gamma distribution with parameters: $E(\delta) = 1$ and $Var(\frac{1}{v})$.

The expected value of y for the Negative binomial distribution is the same as for Poisson distribution but the conditional variance differs:

$$Var(y_i/x_i) = \lambda_i(1 + \frac{\lambda_i}{v_i}) = exp(x_i\beta)(1 + \frac{exp(x_i\beta)}{v_i}) \tag{3.7}$$

since λ and v are positive, the conditional variance of y must exceed the conditional mean, v is the same for all individuals:

$$v_i = \alpha^{-1} \tag{3.8}$$

for $\alpha > 0$, α is the dispersion parameter since increasing α increases the conditional variance of y.

$$Var(y_i/x_i) = \lambda_i(1 + \frac{\lambda_i}{\alpha^{-1}}) = exp(x_i\beta)(1 + \frac{exp(x_i\beta)}{\alpha^{-1}}) = \lambda_i(1 + \alpha\lambda_i) = \lambda_i + \alpha\lambda_i^2 \tag{3.9}$$

If $\alpha = 0$, the mean and variance are equal (Gujarati, 2009).

3.2. Generalized Linear Models (GLMs)

GLMs are a natural generalization of classical linear models that allow the mean of a population to depend on a linear predictor through a (possibly nonlinear) link function. This allows the response probability distribution to be any member of the exponential family (EF) of distributions. A generalized linear model consists of three model components, the random, systematic and link component (McCullagh & Nelder, 1992).

1. Random component. The response Y is independent and has a distribution in the EF, with density or probability function taking the form;

$$f(y; \theta, \phi) = exp \int \frac{[y - \mu(\theta)]}{\phi v(\mu)} d\mu(\theta) + c(y, \phi), \tag{3.10}$$

   where μ is called the natural parameter, $\phi > 0$ is a dispersion parameter, the unknown parameter θ is called the canonical parameter, $\mu = \mu(\theta) = E(Y)$ and $V(Y) = \phi V(\mu)$, for a given variance function V and known bivariate function c. The EF is very flexible and can model continuous, binary, or count data.

2. Systematic component. For a random sample $Y_1, ..., Y_n$, the linear component is defined as;

$$\eta_i = X_i'\beta, i = 1, ..., n, \tag{3.11}$$

   for some vector of parameters $\beta = (\beta_1, ..., \beta_p)'$ are p unknown regression parameters to be estimated and covariate $X_i = (x_{i1}, ..., x_{ip})'$ associated with observation Y_i.

3. Parametric Link component. A monotonic differentiable link function g describes how the expected response $\mu_i = E(Y_i)$ is related to the linear predictor η_i

$$g(\mu_i) = \eta_i, i = 1, ..., n \tag{3.12}$$

3.2.1. Parametric Link Functions

The parametric link function is the third component of a generalized linear model and relates the linear predictor $\eta = X^{'}\beta$ to μ, the expected value of y. In linear models the mean and the linear predictor are identical, so for normal distributed responses mostly the identity function is chosen as a link, $g(\mu_i) = \mu_i$. In the Poisson case however we are dealing with counts and must therefore have $\mu > 0$, so we can't take the identity link. That's why for Poisson distributed response often $g(\mu_i) = log(\mu_i)$ and $G(\eta_i) = g^{-1}(\eta_i) = exp(\eta_i)$ is taken as the link function, because η may be negative, but μ must not be.

3.3. MLE of the Regression Parameters

For an observed independent random sample $y_1, ..., y_n$, consider the log likelihood of β:

$$l(\beta) = logL(\beta) = \sum_{i=1}^{n} \int \frac{[y_i - \mu_i(\theta)]}{\phi v(\mu_i)} d\mu_i(\theta) + c(y_i, \phi) \tag{3.13}$$

Maximizing the log likelihood function (3.13), we solve for the MLE of the regression parameter as $\hat{\beta}$. Take the derivative of (3.13):

$$\frac{dl(\beta)}{d\beta} = \sum_{i=1}^{n} \frac{dl(\beta)}{d\mu_i} \frac{d\mu_i}{d\beta} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\phi V(\mu_i)} \frac{d\mu_i}{dX_i'\beta} \frac{dX_i'\beta}{d\beta},$$

where

$$\frac{d\mu_i}{dX_i'\beta} = \frac{dg^{-1}(X_i'\beta)}{dX_i'\beta} = \frac{1}{g^{'}(\mu_i)}$$

Hence

$$\frac{dl(\beta)}{d\beta} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\phi V(\mu_i)} \frac{1}{g^{'}(\mu_i)} X_i^{'} \tag{3.14}$$

Note that if Y_i has a normal distribution, then $g^{'}(\mu_i) = 1$, and $V(\mu_i) = 1$ for all i.Setting $\frac{dl(\beta)}{d\beta} = 0$ yields $\sum_{i=1}^{n} X_i(y_i - X_i^{'}\beta) = 0$. In other EF cases, no closed form solution is available to this system of p equations. Instead, to obtain the maximum likelihood estimator (MLE) numerically, we must resort to an iterative algorithm such as Newton Raphson or Fisher scoring methods.

The Newton Raphson method provides successive approximations to the root $\hat{\beta}$ of (3.14). On the r^{th} iteration, the algorithm updates the parameter estimate $\hat{\beta}_r$ with;

$$\hat{\beta}_{r+1} = \hat{\beta}_r - H^{-1}s, r = 1, 2, ...,$$

where H is the Hessian (second derivative) matrix, and s is the gradient (first derivative) vector of the log-likelihood function. Both are evaluated at the current value of the parameter estimate and are given by;

$$s = \sum_i \frac{w_i(y_i - \mu_i)x_i}{V(\mu_i)g^{'}(\mu_i)\phi},$$

$$H = -X^{'}W_0X,$$

where X is the design matrix, x_i is the transpose of the i^{th} row of X, and V is the variance function. The matrix W_0 is diagonal with its i^{th} diagonal element equal to;

$$w_{oi} = w_{ei} + w_i(y_i - \mu_i)\frac{V(\mu_i)g^{''}(\mu_i) + V^{'}(\mu_i)g^{'}(\mu_i)}{[V(\mu_i)]^2[g^{'}(\mu_i)]^3\phi}$$

where

$$w_{ei} = \frac{w_i}{\phi V(\mu_i)[g^{'}(\mu_i)]^2}$$

and w_i is a known weight for each observation. If the weight is not specified, then simply put $w_i = 1$ for each observation. The primes denote derivatives of g and V with respect to μ. The negative of H is called the observed information matrix. The expected value of W_0 is a diagonal matrix W_e with diagonal values w_{ei}. If you replace W_0 with W_e, then the negative of H is called the expected information matrix. W_e is the weight matrix for Fisher's scoring method.

3.3.1. Asymptotic properties of the regression MLE

The MLE $\hat{\beta}$ for the GLM parameters has some nice asymptotic properties when n, the number of observations, tends to infinity.

Lemma 1. *1. $\hat{\beta}$ is an asymptotically unbiased and consitent estimator of β.*

2. $V(\hat{\beta}) \longrightarrow \sum = -H^{-1}$, as $n \longrightarrow \infty$. $H = -X^{'}W_0X$ is the Hessian matrix, while $W_0 = diag(w_{01}, ..., w_{0n})$ is a diagonal weight matrix with $i-th$ element ;

$$w_{oi} = \frac{w_i}{\phi V(\mu_i)(g^{'}(\mu_i))^2} + w_i(y_i - \mu_i)\frac{V(\mu_i)g^{''}(\mu_i) + V^{'}(\mu_i)g^{'}(\mu_i)}{(V(\mu_i))^2(g^{'}(\mu_i))^3\phi}$$

for the known weights w_i and covariate matrix $X = (X_1, ..., X_n)^{'}$.

3. $\hat{\beta} \xrightarrow{d} N(\beta, (X^{'}WX)^{-1}\phi)$, i.e. it converges in distribution.

For proof see (Fahrmeir & Kaufmann, 1985)

3.3.2. Estimation of parameters using the Maximum Likelihood Estimation in Poisson regression

Estimation of parameters in Poisson regression relies on maximum likelihood estimation (MLE) method. Maximum likelihood estimation gives an understanding of the values of the regression coefficients that are more likely to have given rise to the data. The maximum likelihood estimation for Poisson regression is discussed in detail below; let Y_i be the mean for the i^th response, for $i = 1, 2, ..., p$. The mean response is assumed to be a function of a set of explanatory variables, $X_1, X_2, ..., X_p$, the notation $\lambda(X_i, \beta)$ is used to denote the function that relates the mean response λ_i and X_i (the values of the explanatory variables for case i) and β (the values of the regression coefficients). Let's consider the Poisson regression model in the form below;

$$\lambda_i = \lambda(X_i, \beta) = e^{X_i,\beta} \tag{3.15}$$

Then, from the Poisson distribution;

$$P(Y; \beta) = \frac{[\lambda(X_i\beta)]^Y e^{-\lambda(X_i\beta)}}{Y!} \tag{3.16}$$

The likelihood function is given as,

$$L(Y;\beta) = \prod_{i=1}^{N} P(Y;\beta) \tag{3.17}$$

$$= \prod_{i=1}^{N} \frac{[\lambda(X_i\beta)]^Y e^{-\lambda(X_i\beta)}}{Y!} \tag{3.18}$$

The next thing to do is taking natural log of the above likelihood function. Then,differentiate the equation with respect to β and equate the equation to zero. The log likelihood function is given as,

$$LogL(Y_i,\beta) = \sum_{i=1}^{N}[Y_i Log\lambda(X_i,\beta)] - \lambda(X_i,\beta) - Log(Y_i!) \tag{3.19}$$

$$\frac{\partial}{\partial\beta}[LogL(Y;\beta)] = 0 \tag{3.20}$$

In this case, no closed form solution is available to this system of p equations. Instead, to obtain the maximum likelihood estimator (MLE) numerically, we must resort to an iterative algorithm such as Newton Raphson or Fisher scoring methods.This procedure will estimate the values of β. Maximum likelihood estimation produces Poisson parameters that are consistent, asymptotically normal and asymptotically efficient (Agresti, 2002).

3.4. Negative Binomial Regression Analysis

The negative binomial regression model is derived by re writing Poisson regression model such that,

$$Log\lambda = \beta_0 + \beta_i X_i + \varepsilon_i \tag{3.21}$$

where e^{ε_i} is a Gamma distributed error-term with mean 1 and variance α^2. This addition allows the variance to differ from the mean as,

$$Var(Y) = \lambda(1+\alpha\lambda) = \lambda + \alpha\lambda^2 \tag{3.22}$$

α also acts as a dispersion parameter. Poisson regression model is regarded as a limiting model of the negative binomial regression model as α approaches zero, which

means that the selection between these two models is dependent upon the value of α. The negative binomial distribution has the form,

$$P(Y=y)=\frac{\Gamma(\frac{1}{\alpha}+y)}{\Gamma(\frac{1}{\alpha})y!}[\frac{\frac{1}{\alpha}}{(\frac{1}{\alpha})+\lambda}]^{\frac{1}{\alpha}}[\frac{\lambda}{(\frac{1}{\alpha})+\lambda}]^{y} \tag{3.23}$$

where $\Gamma(.)$ is a gamma function. This results in the likelihood function,

$$L(Y_i)=\prod_i \frac{\Gamma(\frac{1}{\alpha}+y_i)}{\Gamma(\frac{1}{\alpha})y_i!}[\frac{\frac{1}{\alpha}}{(\frac{1}{\alpha})+\lambda_i}]^{\frac{1}{\alpha}}[\frac{\lambda_i}{(\frac{1}{\alpha})+\lambda_i}]^{y_i} \tag{3.24}$$

Maximum likelihood estimation is used to estimate parameters in negative binomial. In addition, the interpretation of regression coefficients for negative binomial regression is the same as for Poisson regression.

3.4.1. Maximum Likelihood Estimation of Negative Binomial Regression

The regression coefficients are estimated using the method of maximum likelihood. Cameron and Trivedi (2013) gives the logarithm of the likelihood function as;

$$L=\sum_{i=1}^{n}\{log[\Gamma(y_i+\alpha^{-1})]-log[\Gamma(\alpha^{-1})]-log[\Gamma(y_i+1)]-\alpha^{-1}log(1+\alpha\lambda_i)-y_i log(1+\alpha\lambda_i)+y$$

(3.25)

Rearranging gives;

$$L=\sum_{i=1}^{n}\{(\sum_{j=0}^{y_i-1} log(j+\alpha^{-1}))-log(\Gamma(y_i+1))-(y_i+\alpha^{-1})log(1+\alpha\lambda_i)+y_i log(\lambda_i)+y_i log(\alpha)\}$$

(3.26)

The first derivatives of L were given by Cameron and Trivedi (2013) and Lawless (1987) as;

$$\frac{\partial L}{\partial \beta_j}=\sum_{i=1}^{n}\frac{x_{ij}(y_i-\lambda_i)}{1+\alpha\lambda_i}, j=1,2,...,k$$

$$\frac{\partial L}{\partial \alpha}=\sum_{i=1}^{n}\{\alpha^{-2}(log(1+\alpha\lambda_i)-\sum_{j=0}^{y_i-1}\frac{1}{j+\alpha^{-1}})+\frac{y_i-\lambda_i}{\alpha(1+\alpha\lambda_i)}\}$$

$$-\frac{\partial^2 L}{\partial\beta_r\partial\beta_s} = \sum_{i=1}^{n} \frac{\lambda_i(1+\alpha y_i)x_{ir}x_{is}}{(1+\alpha\lambda_i)^2}, r, s = 1, 2, ..., k$$

$$-\frac{\partial^2 L}{\partial\beta_r\partial\alpha} = \sum_{i=1}^{n} \frac{\lambda_i(y_i-\lambda_i)x_{ir}}{(1+\alpha\lambda_i)^2}, r, = 1, 2, ..., k$$

$$-\frac{\partial^2}{\partial\alpha^2} = \sum_{i=1}^{n}\{\sum_{j=0}^{y_i-1}(\frac{j}{1+\alpha j})^2 + 2\alpha^{-3}log(1+\alpha\lambda_i) - \frac{2\alpha^{-2}\lambda_i}{1+\alpha\lambda_i} - \frac{(y_i+\alpha^{-1})\lambda_i^2}{(1+\alpha\lambda_i)^2}\}$$

Equating the gradients to zero gives the following set of likelihood equations;

$$\sum_{i=1}^{n} \frac{x_{ij}(y_i-\lambda_i)}{1+\alpha\lambda_i} = 0, j = 1, 2, ..., k$$

$$\sum_{i=1}^{n}\{\alpha^{-2}(log(1+\alpha\lambda_i) - \sum_{j=0}^{y_i-1}\frac{1}{j+\alpha^{-1}}) + \frac{y_i-\lambda_i}{\alpha(1+\alpha\lambda_i)}\} = 0$$

In this case, no closed form solution is available to this system of equations. Instead, to obtain the maximum likelihood estimator (MLE) numerically, we must resort to an iterative algorithm such as Newton Raphson or Fisher scoring methods.This procedure will estimate the values of β. Maximum likelihood estimation produces Negative Binomial parameters that are consistent, asymptotically normal and asymptotically efficient (Cameron & Trivedi, 2013).

3.4.2. Newton-Raphson Method

Under suitable regularity conditions,the maximum likelihood estimator is a solution to the same equation,

$$S(\theta) = S(X;\theta) = \frac{\partial I(\theta)}{\partial\theta} = \frac{\partial LogL(\theta\theta;X)}{\partial\theta} = 0, \tag{3.27}$$

where $S(\theta) = S(X;\theta)$ is the score statistic. Generally the solution to this equation must be calculated by iterative methods. One of the most common methods is the

Newton-Raphson method and this is based on successive approximations to the solution, using Taylor's theorem to approximate the equation. Thus, we take an initial value θ_0 and write $0 = S(\theta_0) - J(\theta_0)(\theta - \theta_0)$, ignoring the reminder term. Here

$$J(\theta) = J(\theta; X) = -\frac{\partial S(\theta)}{\partial \theta} = -\frac{\partial^2 I(\theta)}{\partial \theta^2} \tag{3.28}$$

Solving this equation for θ then yields a new value θ_1

$\theta_1 = \theta_0 + J(\theta_0)^{-1} S(\theta_0)$

and we keep repeating this procedure as long as $|S(\theta_j)| > \varepsilon$, i.e.

$\theta_{k+1} = \theta_k + J(\theta_k)^{-1} S(\theta_0)$

Clearly, $\hat{\theta}$ is a fixed point of this iteration as $S(\hat{\theta}) = 0$ and, conversely,any fix point is a solution to the likelihood equation. If $\hat{\theta}$ is a local maximum for the likelihood function, we must have

$J(\hat{\theta}) = -\frac{\partial^2 I(\theta)}{\partial \hat{\theta}^2} > 0.$

The quantity $J(\hat{\theta})$ determines the sharpness of the peak in the likelihood function around its maximum. It is also known as the observed information. Occasionally we also use this term for $J(\theta)$ where θ is arbitrary but strictly speaking this can be quite inadequate as $J(\theta)$ may well be negative (although positive in expectation) (Akram & Ann, 2015).

Recall that the (expected) Fisher information is $I(\theta) = E\{J(\theta)\}$ and that for large i.i.d samples it holds approximately that $\hat{\theta} \sim N(\theta, I(\theta)^{-1})$. In contrast to the observed information, $I(\theta)$ is non-negative everywhere, and in regular cases even strictly positive. But it is also approximately true, under the same assumptions that $\sqrt{J(\hat{\theta})(\hat{\theta})} \sim N(0, 1)$, we could write $\hat{\theta} \sim N(\theta, J(\hat{\theta})^{-1})$.

The observed information is in many ways preferable to the expected information. Indeed, $\hat{\theta}$ is approximately sufficient and $J(\hat{\theta})$ is approximately ancillary. The iteration becomes; choose an initial value θ; calculate $S(\theta)$ and $J(\theta)$;

while $|S(\theta)| > \varepsilon$ repeat

1 $\theta \leftarrow \theta + J(\theta)^{-1} S(\theta)$

2 $calculate S(\theta)$ and $J(\theta)$ go to 1

Return θ;

Other criteria for terminating the iteration can be used. To get a criterion which is insensitive to scaling of the variables, one can instead use the criterion $J(\theta)^{-1}S(\theta)^2 > \varepsilon$. Note that, as a by-product of this algorithm, the final value of $J(\theta)$ is the observed information which can be used to assess the uncertainty of $\hat{\theta}$. If θ_0 is chosen sufficiently near $\hat{\theta}$ convergence is very fast. It can be computationally expensive to evaluate $J(\theta)$ a large number of times. This is sometimes remedied by only changing J every 10 iterations or similar.

Another problem with the Newton-Raphson method is its lack of stability. When the initial value θ_0 is far from θ it might wildly oscillate and not converge at all. This is sometimes remedied by making smaller steps as $\theta \leftarrow \theta + \gamma\{J(\theta) + S(\theta)^2\}^{-1}S(\theta)$ as this avoids taking large steps when $S(\theta)$ is large. The iteration has a tendency to be unstable for many reasons, one of them being that $J(\theta)$ may be negative unless θ already is very close to the MLE $\hat{\theta}$. In addition, $J(\theta)$ might sometimes be hard to calculate.

3.4.3. Method of Scoring

Fisher (1922), introduced the method of scoring which simply replaces the observed second derivative with its expectation to yield the iteration

$\theta \leftarrow \theta + I(\theta)^{-1}S(\theta)$.

$I(\theta)$ is easier to calculate and $I(\theta)$ is always positive. This generally stabilizes the algorithm, but here it can also be necessary to iterate as;

$\theta \leftarrow \theta + \gamma\{I(\theta) + S(\theta)^2\}^{-1}S(\theta)$.

In the case of n independent and identically distributed observations we have $I(\theta) = nI_1(\theta)$ so

$\theta \leftarrow \theta + I_1(\theta)^{-1}S(\theta)/n$

where $I_1(\theta)$ is the Fisher information in a single observation. In a linear canonical one-parameter exponential family

$f(X;\theta) = b(X)e^{\theta t(X) - c(\theta)}$

we get

$J(\theta) = \frac{\partial^2}{\partial\theta^2}\{c(\theta) - \theta t(X)\} = c^{''}(\theta) = I(\theta)$.

For canonical exponential families the method of scoring and the method of New-Raphson coincide. If we let $V(\theta) = c^{''}(\theta) = I(\theta) = V(t(X))$ the iteration becomes $\theta \leftarrow \theta + v(\theta)^{-1}S(\theta)/n$.

The identity of Newton-Raphson and the method of scoring only holds for the canonical parameter.

If $\theta = g(\mu)$

$J(\mu) = \frac{\partial^2}{\partial\mu^2}[c\{g(\mu)\} - g(\mu)t(X)]$

$=\frac{\partial}{\partial\mu}[g^{'}(\mu)\tau\{g(\mu)\} - g^{'}(\mu)t(X)]$

$=V\{g(\mu)\}\{g^{'}(\mu)\}^2 + g^{''}(\mu)[\tau\{g(\mu)\} - t(X)]$

where we have let $\tau(\theta) = c^{'}(\theta) = E_\theta t(X)$ and $V(\theta) = c^{''}(\theta) = V_\theta t(X)$.

The method of scoring is simpler because the last term has the expectation equal to 0:

$I(\mu) = E\{J(\mu)\} = V\{g(\mu)\}\{g^{'}(\mu)\}^2$.

3.4.4. The Multi-parameter case

The considerations on the previous over heads readily generalize to the multi-parameter case. The approximation to the score equation becomes;

$0 = S(\theta_0) - J(\theta_0)(\theta - \theta_0)$

where

$S(\theta)_r = \frac{\partial I(\theta)}{\partial\theta_r}, J(\theta)_{rs} = -\frac{\partial^2 I(\theta)}{\partial\theta_r\partial\theta_s}$,

i.e. $S(\theta)$ is the gradient and $-J(\theta)$ the Hessian of $I(\theta)$. The iterative step can still be written as;

$\theta \leftarrow \theta + J(\theta)^{-1}S(\theta)$ where we have to remember that the score statistic S is a vector and the Hessian $-J$ a matrix. The lack of stability of the Newton-Raphson algorithm is not any better in the multi parameter case.

On the contrary, there are not only problems with negativity, but the matrix can be singular and not invertible or it can have both positive and negative eigen values. Recall that a symmetric matrix A is positive definite if all its eigen values are positive or, equivalently, if $X^TAX > 0$ for all $X \neq 0$.

Sylvester's theorem says that:

Theorem 2. *A is positive definite if and only if $det(A_R) > 0$ for all sub matrices A_R of the form $\{a_r s\}, r, s = 1, ..., R$*

.

It is therefore advisable to replace $J(\theta)$ with its expectation, the Fisher information matrix,i.e. Iterate as;

$\theta \leftarrow \theta + I(\theta)^{-1}S(\theta)$

where now $I(\theta)$ is the Fisher information matrix which is always positive definite if the model is not over - parameterized. Also in the multi-parameter case it can be advisable to stabilize additionally, i.e. by iterating as;

$\theta \leftarrow \theta + \gamma\{I(\theta) + S(\theta)S(\theta)^T\}^{-1}S(\theta)$

or

$\theta \leftarrow \theta + \gamma\{I(\theta) + S(\theta)^T S(\theta)E\}^{-1}S(\theta)$

where E is the identity matrix.

In a multi-parameter curved exponential family with densities ;

$f(X;\beta) = b(X)e^{\theta(\beta)^T t(X) - c\{\theta(\beta)\}}$

where β is d- dimensional, we get;

$J(\beta) = \frac{\partial^2}{\partial\beta\partial\beta^T}[c\{\theta(\beta)\} - \theta(\beta)^T t\{X\}]$

$= \frac{\partial}{\partial\beta}[(\frac{\partial\theta}{\partial\beta})^T\tau\{\theta(\beta)\} - (\frac{\partial\theta}{\partial\beta})^T t(X)]$

$= \frac{\partial^2\theta}{\partial\partial^T}[\tau\{\theta(\beta)\} - t(X)] + (\frac{\partial\theta}{\partial\beta})^T V\{\theta(\beta)\}(\frac{\partial\theta}{\partial\beta})$,

where the first term has expectation zero so,

$I(\beta) = E\{J(\theta)\} = (\frac{\partial\theta}{\partial\beta})^T V\{\theta(\beta)\}(\frac{\partial\theta}{\partial\beta})$

In the multi-parameter case it is in wide generality approximately true that;

$\hat{\theta} \sim N_d(\theta, I(\theta)^{-1})$ or with a slight imprecision ;

$\hat{\theta} \sim N_d(\theta, J(\hat{\theta})^{-1})$

where N_d is the d-dimensional Gaussian distribution. In particular it holds approximately that; $(\hat{\theta} - \theta)^T I(\theta)(\hat{\theta} - \theta)$ $(\hat{\theta} - \theta)^T I(\hat{\theta})(\hat{\theta} - \theta)$ $(\hat{\theta} - \theta)^T J(\hat{\theta})(\hat{\theta} - \theta)$ $\chi^2_{(d)}$.

3.5. Establishing Appropriate Statistical Models

In this study, data on monthly Malaria incidence in Apac District for the period January 2007 to December 2016 were obtained from the Ministry of health, Uganda. Climate data were obtained from Uganda National Meteorological Authority. The response variable is the malaria incidence where as the climate variables are the explanatory variables. In this study, the association between malaria incidences and climate variables was modeled using Poisson and Negative Binomial Regression models respectively. We are particularly interested in the significance of rainfall and temperature on the malaria incidences. This knowledge is important to the development of malaria warning systems in Apac District, Northern Uganda and hence enable effective malaria control measures to be put in place in a timely. The aim of this work was also to develop a predictive model that can forecast the incidence of malaria incidences using the reported cases, temperature and rainfall.

The generalized linear models were applied to fit the malaria incidence data as a function of rainfall and average temperature.

3.6. Goodness of Fit

Deviance was used to test the goodness of fit of the model. Deviance is a measure of discrepancy between observed and fitted values. According to likelihood ratio (LR) theory, under regularity condition and asymptotically;

$$\frac{D(y;\hat{\mu})}{\phi} = 2(log(\theta) - log(\hat{\theta}) = -2log(\lambda(y)) \sim \chi^2_{n-p}$$

if the model represented by $\hat{\theta}$ is an adequate model with p parameters, where $D(y;\hat{\mu})$ is termed the deviance of the fitted model that defines $\hat{\mu}$. ϕ is the dispersion parameter. In expectation, if the fitted model is adequate, $\frac{D(y;\hat{\mu})}{\phi} \approx n-p$ which defines a heuristic model adequacy assessment (i.e if $\frac{D(y;\hat{\mu})}{\phi} \approx n-p$, then the model can be considered adequate). The deviance for Poisson responses takes the form;

$$D = 2\sum_{i=1}^{n}\{y_i log(\frac{y_i}{\hat{\mu}_i}) - (y_i - \hat{\mu}_i)\}$$

The first term represents 'twice a sum of observed times log of observed over fitted'. The second term, a sum of differences between observed and fitted values, is usually zero, because maximum likelihood estimations in Poisson models have the property of reproducing marginal totals. For large samples of the distribution, the deviance is approximately a chi-square with $n - p$ degrees of freedom, where n is the number of observations and p the number of parameters. Therefore, the deviance can be used directly to test the goodness of fit of the model.

In the case of negative binomial regression, the deviance is a generalization of the sum of squares. The maximum possible log likelihood is computed by replacing μ_i with y_i in the likelihood formula. Thus, we have

$$D = 2[L(y_i) - L(\mu_i)] = 2\sum_{i=1}^{n}\{y_i log\frac{y_i}{\mu_i} - (y_i + \alpha^{-1})log\frac{1+\alpha y_i}{1+\alpha\mu_i}\}$$

3.7. Residual analysis

Residual analysis was performed to determine the fit of the models developed. The Poisson regression is a non-normal regression, that is residuals are far from being normally distributed and the variances are non constant. Therefore we assess the model based on quantile residuals which removes the pattern in discrete data by adding the smallest amount of randomization necessary on cumulative probability scale. The quantile residuals are obtained by inverting the distribution function for each response.

Mathematically, let $a_i = lim_{y\uparrow y_i}F(y;\hat{\mu},\hat{\Theta})$ and $b_i = F(y_i;\hat{\mu},\hat{\Theta})$ where F is the cumulative function of the probability density function $f(y;\mu,\Theta)$ then the randomized quantile residuals for y_i is $r_{q,r} = \Phi^{-1}(u_i)$ with u_i the uniform random variable on $(a_i, b_i]$. The randomized quantile residuals are distributed normally barring the variability in $\hat{\mu}$ and $\hat{\Theta}$ (Dunn & Smyth, 1996).

3.8. Pearson correlation

Pearson correlation was used to measure the relationship between expected malaria incidences and rainfall as well as average temperature. Correlation between sets of data

is a measure of how well they are related. The most common measure of correlation in statistics is the Pearson Correlation. The pearson correlation coefficient formula is given below;

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

It's value is between $-1$and 1. It is very rarely see 0, -1 or 1. The values are usually somewhere in between those values. The closer the value of r gets to zero, the greater the variation the data points are around the line of best fit. High correlation is from 0.5 to 1.0 or -0.5 to -1.0, medium correlation is from 0.3 to 0.5 or -0.3 to -0.5 and low correlation is from 0.1 to 0.3 or -0.1 to -0.3 (Mukaka, 2012).

3.9. Multicollinearity Test

A variance inflation factor(VIF) was used to detect multicollinearity in regression analysis. Multicollinearity is when theres correlation between predictors (i.e. independent variables) in a model; its presence can adversely affect your regression results. The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

$$VIF = \frac{1}{1 - R_i^2} \tag{3.29}$$

where R_i^2 is the coefficient of determination. Variance inflation factors range from 1 upwards. The numerical value for VIF tells you (in decimal form) what percentage the variance (i.e. the standard error squared) is inflated for each coefficient. For example, a VIF of 1.9 tells you that the variance of a particular coefficient is 90% bigger than what you would expect if there was no multicollinearity if there was no correlation with other predictors. A rule of thumb for interpreting the variance inflation factor: 1 = not correlated, between 1 and 5 = moderately correlated, greater than 5 = highly correlated. Exactly how large a VIF has to be before it causes issues is a subject of debate. In general, a VIF above 10 indicates high correlation and is cause for concern (Vatcheva et al., 2016).

3.10. Methods used in Forecasting

3.10.1. ARIMA Models

In this study R statistical software was used to develop ARIMA (Autoregressive integrated moving average) models for forecasting. ARIMA model was analyzed with the application of Box-Jenkins approach in which the data was analyzed and used to identify, estimate and select the best model. First and foremost we checked the data for stationarity before using it to develop ARIMA models. Augmented Dickey-Fuller test was used to test the null hypothesis that the data is non-stationary versus the alternative hypothesis that the data is stationary. When the data is found to be non stationary, it is differenced to make it stationary (McLeod & Li, 1983). When stationarity is obtained with a differenced ARIMA parameter d (the number of times the series is differenced to achieve stationarity), we then identified the order of the two processes that construct the ARIMA model that is the AR and MA. After which we estimated the parameters of the models.

In order to select an appropriate subclass of models from the general ARIMA(p,d,q), the following approaches of the ARIMA model were used to develop a model to forecast malaria incidences from historical malaria incidence data in Apac District. The Autoregressive integrated moving average (ARIMA) models or Box-Jenkins methodology, are a class of linear models that use historical values of a single variable to forecast it's future values; hence they are classified as univariate methods. The model can represent both stationary and non stationary time series. However, for adequate ARIMA modeling, a time series should be stationary with respect to mean and variance (Makridakis et al., 2008). To obtain stationary time series, the original time series should be transformed, such as a log transformation, time series differencing or variance stabilization. Once a stationary series has been obtained, then a satisfactory model has been obtained and can be used to forecast expected number of cases for a given number of future time intervals.

Consider a discrete time series of equally spaced n observations in time:

$$Y_t = Y_1, Y_2, ..., Y_{n-1}, Y_n \tag{3.30}$$

An equation of ARIMA model is combining two processes; the autoregressive process (AR) which expresses Y_t as a function of it's past values and the moving average process (MA) as a function of past values of the error term. There we represent the ARIMA process as;

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} - e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - ... - \theta_q e_{t-q} \quad (3.31)$$

where $\phi' s$ and $\theta' s$ are the coefficients of the AR and MA processes respectively, and p and q are the number of past values of Y_t and the error terms respectively.

The general notation of the ARIMA models is ARIMA (p, d, q) where p is the order of the autoregressive component, d is the order of differencing used and q is the order of the moving average component in the model. Differencing a series involves subtracting it's current and previous values d times. Often differencing is used to stabilize the series when the stationarity assumption is not met. ARIMA models can also be specified through a seasonal structure. In this case, the model is specified by two sets of order parameters: $(p, d, q)(P, D, Q)^s$ where p and P- are the autoregressive and seasonal autoregressive respectively, d and D- are the non-seasonal differences and seasonal differencing respectively, q and Q- are the moving average parameters and seasonal moving average parameters respectively, and s represents the length of the seasonal period.

An autoregressive component, AR(p) refers to the use of past values in the regression equation for the series Y. The autoregressive parameter p specifies the number of lags used in the model. AR(p) is represented as;

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + e_t \quad (3.32)$$

where $\phi_i^{'} s, i = 1, 2, ..., p$ are the model parameters. Usually ARMA models are manipulated using the lag operator notation. The lag or backshift operator is defined as $Ly_t = y_{t1}$. Polynomials of lag operator or lag polynomials are used to represent ARMA models as follows ; AR(p) model $_t = \phi(L) y_t$. It is shown that an important property of AR(p) process is invertibility, i.e. an AR(p) process can always be written in terms of an MA(∞) process. Whereas for an MA(q) process to be invertible, all the

roots of the equation $\theta(L) = 0$ must lie outside the unit circle. This condition is known as the Invertibility Condition for an MA process.

A moving average component MA(q) represents the error of the model as a combination of previous error terms e_t. The order q determines the number of terms to be included in the model. MA(q) is represented as;

$$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - ... - \theta_q e_{t-q} \tag{3.33}$$

where $\theta_i^{'} s, i = 1, 2, ..., q$ are the model parameters. The random shocks are assumed to be a white noise process, that is a sequence of independent and identically distributed (i.i.d) random variables with zero mean and a constant variance σ^2 . Generally, the random shocks are assumed to follow the typical normal distribution. Thus conceptually a moving average model is a linear regression of the current observation of the time series against the random shocks of one or more prior observations.

After the ARIMA models for malaria incidences data and log transformed malaria data were developed, the accuracy of these models were tested and compared in order to choose the best predictive models. To select the best model we used Akaike information criterion (AIC), Bayesian information criterion to estimate the quality of each model relative to each other and the Mean Absolute Error (MAE) to measure the average magnitude of errors in the models. Taking all these measures into account, the best model chosen was one with lower AIC, BIC and MAE values.

3.11. Accuracy Tests

3.11.1. Akaike Information Criterion

The Akaike information criterion (AIC) is a measure of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Hence, AIC provides a means for model selection (Kihoro et al., 2004). Let L be the maximized value of the likelihood function for the model; let k be the number of estimated parameters in the model. Then the AIC value of the model is given by:

$$AIC = 2k - 2\ln L \tag{3.34}$$

Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value. Hence AIC rewards goodness of fit (as assessed by the likelihood function), but it also includes a penalty that is an increasing function of the number of estimated parameters. The penalty discourages over fitting (increasing the number of parameters in the model almost always improves the goodness of the fit).

3.11.2. Bayesian Information Criterion (BIC)

The Bayesian information criterion (BIC) is a criterion for model selection among a finite set of models (Kihoro et al., 2004). The model with the lowest BIC is preferred. It is based on the likelihood function and it is closely related to the Akaike information criterion (AIC). AIC and BIC feature the same goodness-of-fit. When fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in over fitting. Both BIC and AIC resolve this problem by introducing a penalty term for the number of parameters in the model. The penalty term is larger in BIC than in AIC. The BIC value of the model is given by:

$$BIC = k\ln n - 2\ln L \tag{3.35}$$

where n is the number of observations or the sample size.

3.11.3. Mean Absolute Error (MAE)

The simplest measure of forecast accuracy is called Mean Absolute Error (MAE). MAE is simply, as the name suggests, the mean of the absolute errors. The absolute error is the absolute value of the difference between the forecasted value and the actual value. MAE tells us how big of an error we can expect from the forecast on

average. The mean absolute error is defined as (Park, 1999);

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|e_t|$$

Its properties are; It measures the average absolute deviation of forecasted values from original ones, it is also termed as the Mean Absolute Deviation (MAD), it shows the magnitude of overall error, occurred due to forecasting, the effects of positive and negative errors do not cancel out, MAE does not provide any idea about the direction of errors, for a good forecast, the obtained MAE should be as small as possible, MAE depends on the scale of measurement and data transformations, extreme forecast errors are not panelized by MAE.

3.12. Unit roots

In empirical analysis using time series data, it is essential to establish the presence or absence of unit root in the series being studied. The presence or absence of unit roots helps to identify the nature of the processes that generates the time series data and to investigate the order of integration of a series. This is because, contemporary econometrics has indicated that, regression analysis using non-stationary time series variables produce spurious regression since standard results of OLS do not hold. A variable is said to be stationary if the mean, the variance and the covariance of the series are finite and are time invariant. Where,

$$E(Y_t) = E(Y_{t-1}) = \mu$$

which is a constant and

$$cov(Y_t, Y_{t-l}) = \gamma_l$$

which depends only on the lag l and not on time t. If there exist no unit root, the time series fluctuates around a constant long-run mean with finite variance which does not depend on time. There are several proposed quantitative methods of testing for stationarity of a time series variable. This study however employed graphical approaches of

the time series plots i.e., Autocorrelation functions (ACF) and Partial Autocorrelation functions (PACF). In graphical form, a time series plot which do not show mean reversion gives an indication that the levels of the series are non-stationary. Also a slow decaying ACF plot also gives an indication of non-stationarity of a time series. The qualitative method used in this research is the Augmented Dickey-Fuller (ADF) test.

3.12.1. Augmented Dickey Fuller (ADF) Unit Root Test

This study employed the Augmented Dickey-Fuller (ADF) test to determine the stationarity of the malaria time series data. The ADF test proposed by Dickey and Fuller (1979) is an upgraded form of the Dickey-Fuller (DF) test. This test is based on the assumption that the series follow a random walk with model;

$$R_t = \Phi Y_{t-1} + u_t \tag{3.36}$$

and tests the hypothesis: $H_0 : \Phi = 1$ (Non-stationary) against $H_1 : \Phi < 1$ (Stationary) where ϕ is the characteristic root of an AR polynomial and u_t is an uncorrelated white noise series with zero mean and constant variance σ^2. When $\Phi = 1$, equation (3.36) does not satisfy the weakly stationary condition of an AR (1) model hence the series becomes a random walk model known as a unit root non-stationary time series. Subtracting Y_{t-1} from both sides of equation (3.36) we get;

$$\Delta R_t = \varphi Y_{t-1} + u_t \, , t = (1, ..., T) \tag{3.37}$$

where $\varphi = \Phi - 1$ and $\Delta R_t = Y_t - Y_{t-1}$. For estimating the existence of unit roots using equation (3.37), we test hypothesis $H_0 : \varphi = 0$ against $H_1 : \varphi \neq 0$. Under H_0 , if $\varphi = 0$, then $\Phi = 1$, thus the series has a unit root hence is non-stationary. The rejection or otherwise of the null hypothesis, H_0 is based on the t-statistic critical values of the Dickey Fuller statistic. The Dickey Fuller test assumes that the error terms are serially uncorrelated, however, the errors terms of the Dickey Fuller test do show evidence of serial correlation. Therefore, the proposed ADF test includes the lags of the first difference series in the regression equation to make u_t a white noise. The Dickey and

Fuller (1979) new regression equation is given by;

$$\Delta R_t = \varphi Y_{t-1} + \sum_{j=1}^{p} \gamma_j \Delta r_{t-j} + u_t \, , t = (1, ..., T) \tag{3.38}$$

If the intercept and time trend $\beta + \alpha t$ are included, then equation (3.38) is written as;

$$\Delta R_t = \beta + \alpha t + \varphi Y_{t-1} + \sum_{j=1}^{p} \gamma_j \Delta r_{t-j} + u_t \, , t = (1, ..., T) \tag{3.39}$$

where β is an intercept, α defines the coefficient of the time trend factor, $\sum_{j=1}^{p} \gamma_j \Delta r_{t-j}$ defines the sum of the lagged values of the response variable ΔR_t and p is the order of the autoregressive process. If φ of the Augmented Dickey Fuller model is zero , then there exist a unit root in the time series variable considered, hence the series is not covariance stationary. The choice of the starting augmentation order depends on the periodicity of the data, the significance of γ_i estimates and the white noise residuals series u_t. The ADF test statistic is given by;

$$F_\tau = \frac{\hat{\varphi}}{SE(\hat{\varphi})} \tag{3.40}$$

where $\hat{\varphi}$ is the estimate of φ and $SE(\hat{\varphi})$ is the standard error of the least square estimate of $\hat{\varphi}$. The null hypothesis H_0 is rejected if, the $p - value < \alpha$(significance level). If the series is not stationary, it is transformed by differencing to make it stationary and stationarity tested again. If the time series is not stationary but its first difference is stationary, then the series is said to be an integrated process of order one (1) or simply an $I(1)$ process.

3.12.2. Univariate Ljung-Box Test

The study employed the univariate Ljung and Box (1978) test to test jointly whether or not several autocorrelations r_l of the residuals of the individual ARIMA models fitted were zero. It is based on the assumption that the residuals contain no serial correlation (no autocorrelation) up to a given lag m. The univariate Ljung-Box statistic is given by:

$$Q(m) = T(T+2) \sum_{l=1}^{m} \frac{r_l^2}{T-l} \tag{3.41}$$

where r_l represents the residual sample autocorrelation at lag l, T is the size of the series, m is the number of time lags included in the test. $Q(m)$ has an approximately chi-square distribution with m degrees of freedom. We fail to reject H_0 and conclude at α-level of significance that, the residuals are free from serial correlation when the $pvalue$ is greater than the significance level.

3.12.3. Univariate ARCH-LM Test

For a fitted model to adequately fit a series, the variance of the models' residuals must be constant over time. The univariate ARCH-LM test proposed by Engle (1982) was used in this research to check for the presence or absence of conditional heteroscedasticity in the residuals of the individual equations of the model fitted. If there exist no ARCH-effect, it implies that the residuals of the model are homoscedastic and have constant variance. This statistic uses the linear regression model;

$$u_t^2 = a_0 + a_1 u_{t-1}^2 + ... + a_m u_{t-m}^2 + e_t,\ t = m+1, ..., T \tag{3.42}$$

where e_t is the error term, T is the sample size and m is a positive integer. The ARCH-LM statistic tests the hypothesis that; $H_0 = a_1 = ... = a_m = 0$ no ARCH effect against

$H_1 = a_1 \neq ... \neq a_m \neq 0$ ARCH effect exist

The ARCH-LM test statistic is calculated as;

$$LM = TR^2 \tag{3.43}$$

where R^2 is the coefficient of determination for the auxiliary regression. The decision rule is to reject H_0 and conclude that there is conditional heteroscedasticity (ARCH-effect) in the residuals of the model if $LM > \chi_m^2$, or if the $P - value < \alpha$, where m is the lag order of ARCH-effect and α is the significance level chosen.

3.12.4. Jarque Bera test

A Jarque-Bera test was performed to test for normality. Description:

The Jarque-Bera test (Jarque & Bera, 1987) is based on the sample skewness and

sample kurtosis. The Jarque-Bera test statistic is defined as:

$$\frac{N}{6}(S^2 + \frac{(K-3)^2}{4}) \tag{3.44}$$

with S, K, and N denoting the sample skewness, the sample kurtosis, and the sample size, respectively. Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point. Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution.

$$S = \frac{E(X-\mu)^3}{\sigma^3} \tag{3.45}$$

$$K = \frac{E(X-\mu)^4}{\sigma^4} \tag{3.46}$$

4. Results And Discussions

4.1. Data Exploration

Monthly malaria incidences for the period $2007 - 2016$ was used. The data was obtained from the Ministry of Health, Uganda. Table 4.1 shows results for exploratory analysis for Malaria incidences for the period $2007 - 2016$.

Table 4.1.: Summary of Malaria Incidences

Malaria incidences	Mean	Standard deviation	Range
	9746	3017.864	$5034 - 19289$

From Table 4.1, the minimum value for malaria incidences is 5034, the maximum value is 19289 cases, the mean value for the whole period is 9746 and the standard deviation is 3017.864.

From Figure 4.1, it is observed that the histogram is skewed to the right which is representation of count data. We conclude that malaria incidences is count data and we can model it using a distribution suited for count data which is the Poisson distribution.

4.2. Significance of Rainfall and Temperature on Malaria incidences

The expected malaria incidences was modeled using Poisson regression and the results are presented in Table 4.2. The model examines the association between monthly

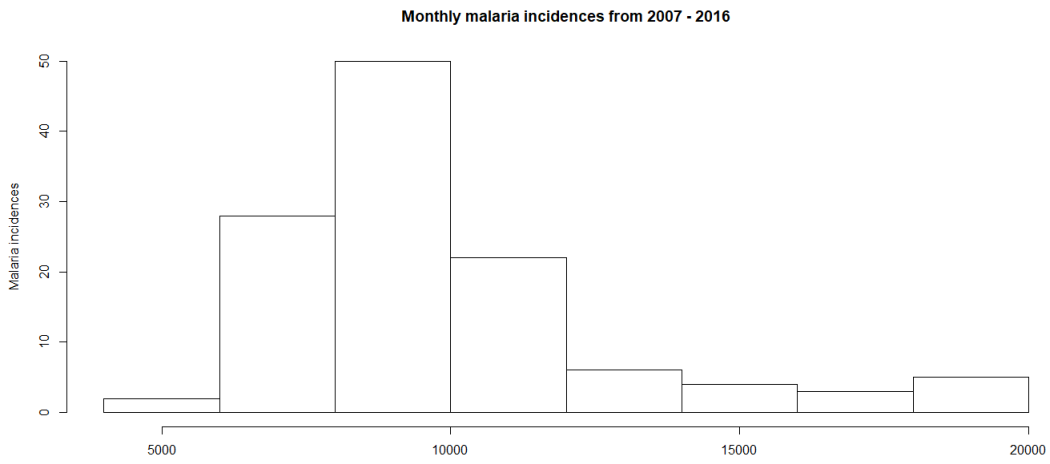

Figure 4.1.: Monthly Malaria incidences over the period 2007-2016

expected malaria incidences with monthly rainfall and monthly average temperature. From the Table 4.2 , it was observed that for every unit increase in rainfall, the expected malaria incidences increases by $e^{0.0007877} = 1.0007880$ and for a unit increase in average temperature, the expected malaria incidences decrease by $e^{-0.03265} = 0.9678773$ obtained from Equation (4.1). Based on the P- values, the average temperature and rainfall significantly affect the expected malaria incidences.

Table 4.2.: Parameter Estimates of Poisson Regression

	Estimate	Standard Errors	P-value
Intercept	9.888	0.03087	0.000000
Rainfall	0.0007877	0.00001493	0.000000
Average Temperature	-0.03265	0.001187	0.000000

The residual deviance for the fitted Poisson regression was given as 89489 on 117 degrees of freedom.

The fitted Poisson model is given as from Table 4.2

$$LogA = 9.8881 + (7.877e^{-04})R - (3.265e^{-02})T = 9.888 + 0.0007877R - 0.03265T \tag{4.1}$$

where A is the expected malaria incidences, R stands for rainfall and T stands for average temperature.

To check the fit of the fitted Poisson model, the value of the residual deviance 89489 on 117 degrees of freedom was considered as observed in Table 4.2, it was observed to be far greater than the number of degrees of freedom. This implies that the ratio $\frac{89489}{117} = 764.863$ which is the dispersion parameter. This value 764.863 is far greater than one. Therefore it can be concluded that the model has lack of fit. If the mean and variance were equal, the residual deviance should be approximately equal to the df for error. The assumption of mean equal to variance of the Poisson random variable hence was violated since the dispersion parameter was not approximately equal to 1, an indication of over dispersion in the data. This meant that the parameters of the model had been over estimated and the standard errors had been under estimated which did not give a true reflection of the model that could provide appropriate expected malaria incidences from 2007 to 2016. The fitted Poisson model had an AIC value of 90813 and a null deviance of 100505 on 119 degrees of freedom.

Table 4.3.: Parameter Estimates for Negative Binomial Regression Model for Rainfall and Average Temperature

	Estimate	Standard Errors	P-value
Intercept	10.1940861	0.7788004	0.000000
Rainfall	0.0008147	0.0003849	0.0343
Average Temperature	-0.0451456	0.0298939	0.1310

The Negative Binomial regression model whose results were presented in Table 4.3, showed null deviance to be 140.31 on 119 degrees of freedom, residual deviance to be 121.36 on 117 degrees of freedom and AIC to be 2225.3.

To address this error, Negative Binomial Regression was used to modify the model so that the case of over dispersion in the data was taken care of and the results were presented in Table 4.3 . It was observed that the Negative Binomial was actually the best model which fit the expected malaria incidences because the dispersion parameter

given by Poisson Regression Model had been reduced from 770 to 1.03. Tests for multicollinearity indicated that a slightly high level of multicollinearity was present (VIF = 11.46 for average temperature and 10.22 for rainfall)

Figures, 4.2, 4.3, 4.4 and 4.5, show plots of the deviance residuals against the normal quantiles based on Poisson model and Negative binomial models respectively. Figure 4.2 for Poisson regression, the plot was not approximately linear just as for Figure 4.4 and Figure 4.5. This indicated poor fit of the models. Figure 4.3, for the Negative Binomial model relating Malaria incidences and rainfall, the plot was approximately linear. This gave the best fit compared to the rest of the plots.

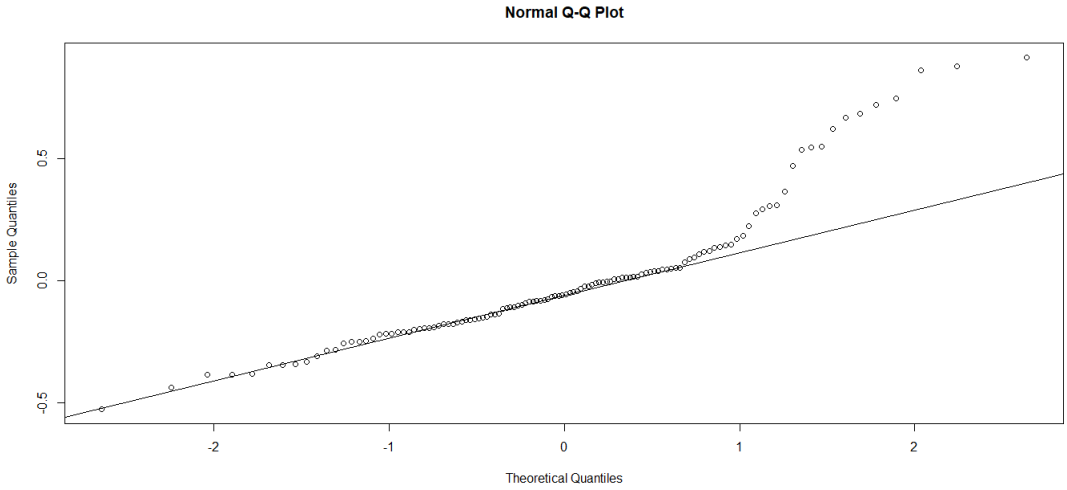

Figure 4.2.: Normal Q-Q plot for Poisson Regression

From the Figure 4.2, the points form a curve instead of a straight line. Normal Q-Q plots that look like this usually imply the model has a lack of fit.

Three models for Negative Binomial regression were considered and compared using the Akaike information criterion (AIC) . The results for the two Negative Binomial Regression models without collinearity respectively are given in Table 4.4 and Table 4.5, respectively. The model from Table 4.5 gave the lowest AIC value that is 2225.4 compared to the other model from Table 4.4 which gave higher AIC value that is 2227.7. Hence Negative Binomial model for the relationship between rainfall and expected malaria incidences whose results were presented in Table 4.5 was considered the model with better fit since it had lower AIC value. The lower the AIC value, the

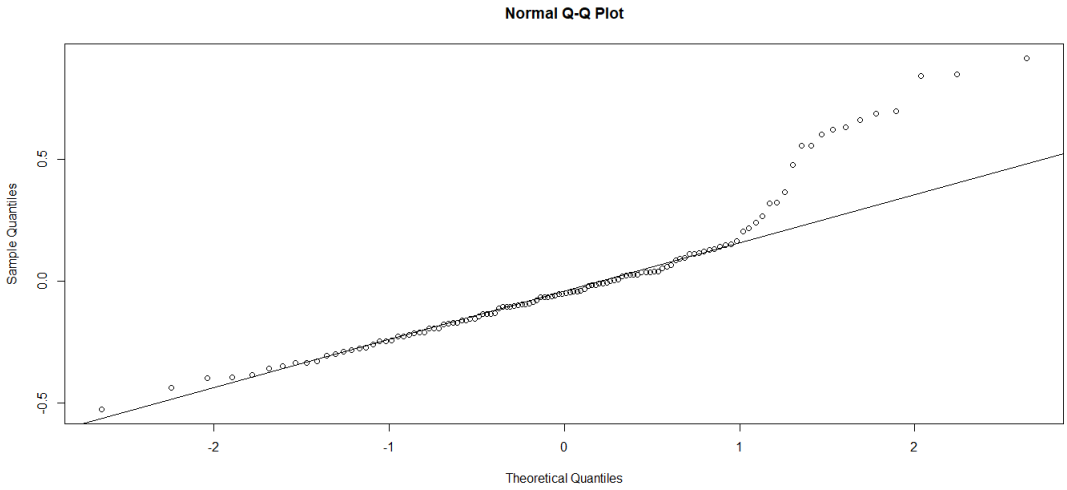

Figure 4.3.: Normal Q-Q plot for Negative Binomial Regression Model between Rainfall and Expected malaria incidences

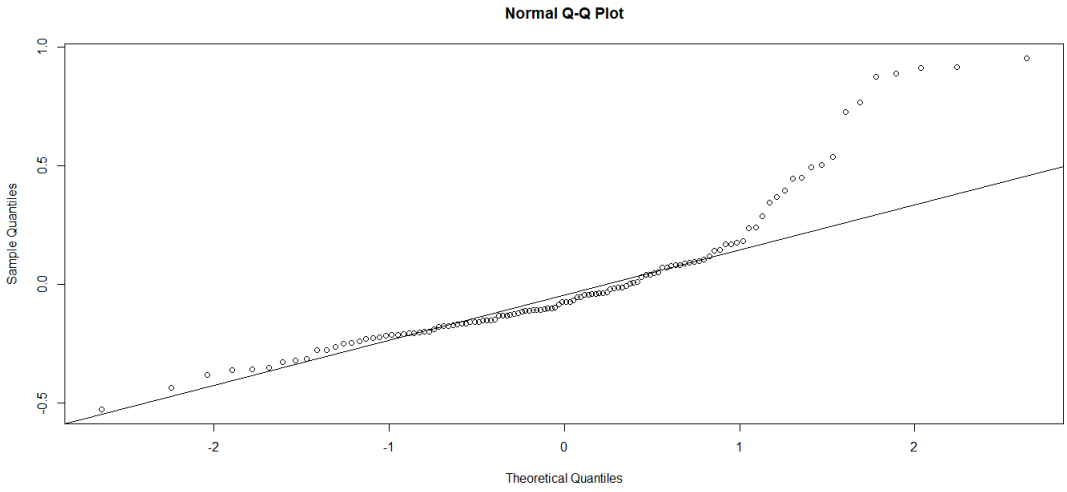

Figure 4.4.: Normal Q-Q plot for Negative Binomial Regression Model between Average Temperature and Expected malaria incidences

better the model.

The Negative Binomial regression model whose results were presented in Table 4.4, showed null deviance to be 135.34 on 119 degrees of freedom, residual deviance to be 121.41 on 118 degrees of freedom and AIC to be 2227.7

The Negative Binomial regression model whose results were presented in Table 4.5

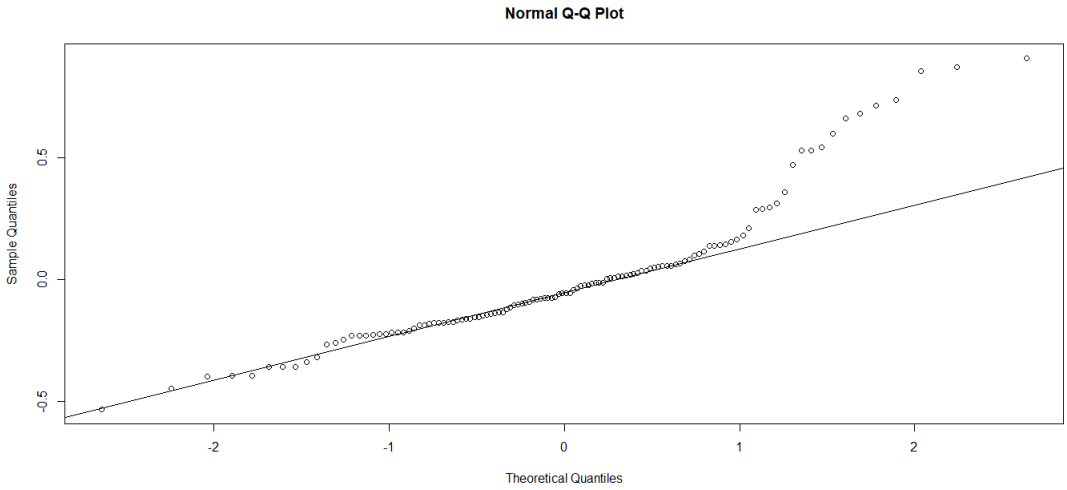

Figure 4.5.: Normal Q-Q plot for Negative Binomial Regression Model between Average Temperature, Rainfall and Expected malaria incidences

Table 4.4.: Parameter Estimates for Negative Binomial Regression Model for Average Temperature

	Estimate	Standard Errors	P-value
Intercept	11.37160	0.54154	0.000000
Average Temperature	−0.08822	0.02178	0.0000512

Table 4.5.: Parameter Estimates for Negative Binomial Regression Model for Rainfall

	Estimate	Standard Errors	P-value
Intercept	9.0252866	0.0436207	0.000000
Rainfall	0.0011832	0.0002779	0.0000206

showed null deviance to be 137.88 on 119 degrees of freedom, residual deviance to be 121.38 on 118 degrees of freedom and AIC to be 2225.4.

The Model whose results were presented in Table 4.3 was not considered because of a negatively strong correlation between the explanatory variables that is rainfall and

average temperature which was reported in Table 4.6 as -0.6985212. Therefore one of the independent variables was removed from the model and two Negative Binomial regression models were developed and results for the parameter estimates presented in Table 4.4 and Table 4.5 respectively. Based on P-value, the results in Table 4.4 and 4.5 showed that average temperature and rainfall significantly affected the expected malaria incidences respectively.

Table 4.6.: Correlation between rainfall and average temperature

Climate Variables	Correlation value
Rainfall and Average Temperature	-0.6985212

Table 4.7.: Pearson correlation test results

	Malaria and Average temperature	Malaria and rainfall
t-value	-3.0311	3.4594
P-value	0.9985	0.0003771
confidence interval	$[-0.4033168, 1.0000000]$	$[0.1598574, 1.0000000]$
degrees of freedom	118	118

Pearson correlation test was performed and the results was presented in Table 4.7. The results show that malaria and rainfall were strongly positively associated based on the P-value which is significant since it was found to be less than 5% significance level and a positive confidence interval. Malaria and Average temperature were found to be negatively associated and not significant given the P-value was greater than 5% significance level.

4.2.1. Interpretation of coefficients

From Table 4.5, we observe that rainfall is very significant at 5% significance level with their significance value equal to 0.01388. For every one unit increase in amount

of rainfall, the expected malaria incidences increases by $e^{0.0011832} = 1.0011839003$ times. From Table 4.3, we observe that rainfall is slightly significant with significance value 0.0343 at 5% significance level while temperature is not significant at 5% significance level with P-value of 0.1310. This indicated the presence of collinearity between rainfall and average temperature. The positive coefficient of rainfall implies that as rainfall increases, the expected malaria incidences also increase.

4.2.2. Discussion on significance of Rainfall and Temperature on Malaria incidences

Malaria is transmitted by the female Anopheles mosquito. The female Anopheles mosquito go through four stages in their life cycle that is egg, larva pupa and adult (Wardrop et al., 2013). The first three stages are aquatic and also depend on the temperature. The adult stage is when the female Anopheles mosquito acts as malaria vector(Wardrop et al., 2013). Once adult mosquitoes have emerged, the temperature, humidity and rainfall determine their chances of survival. To transmit malaria successfully, female Anopheles must survive long enough after they have become infected to allow the parasites they harbour to complete their growth cycle (Kakchapati & Ardkaew, 2011). A conducive climatic environment will also shorten the time required for the parasite development in the mosquito (Agusto et al., 2012). The climate variables can affect the malaria incidences by affecting the life cycle of the mosquito development and the parasite in the mosquitoes.

Pearson correlation between rainfall and average temperature showed a strong negative correlation. This highlights the importance of removing one of the climate variables from the model to avoid invalid association due to collinearity. In this study, rainfall was the only climate variable considered in the Negative Binomial Regression model since it presented the best fit. Negative Binomial regression model relating expected malaria incidences, rainfall and temperature was not selected as the final model due to high correlation between rainfall and average temperature which affected the significance of individual climate variables to expected malaria incidences. The model results showed that average temperature was not significant in the model

while rainfall was weakly significant. This result was seen to contradict the biology of mosquito development.The model relating malaria incidences and average temperature showed a significant positive relationship though it was not the model selected since it had a higher AIC value. Modeling has shown that optimal malaria transmission occurs at 25°C and malaria transmission decreases at temperature above 28°C (Musa et al., 2012). Temperatures below 16°C are also detrimental for survival of mosquitoes (WHO, 2013). Results from previous study (Gomez-Elipe et al., 2007), showed a strong positive association between malaria incidence in a given month and the minimum temperature of the previous month.

In previous studies of climatic effects on malaria incidence, different results on the effect of rainfall on malaria incidence were found. Hove-Musekwa et al. (2008), found rainfall to be significant when precipitation was 2.4 times higher than the normal level. Rainfall plays an important role in the survival of mosquitoes, since water pools from the rain provide a habitat for mosquito larvae to develop. Bloland et al. (1999), found there was no significant effect when rainfall was less than 100mm per month in Yunnan, China. A study in Ethiopia found that rainfall had a significant effect on malaria incidence in hot districts with an altitude lower than $1,650$mm, but not in cold districts with an altitude higher than $1,650$mm (WHO, 2015).

In the present study, there was a positive significant effect between rainfall and malaria incidences, similar to previous findings (WHO, 2015; Yang & Ferreira, 2000). The positive effects were reasonable because rain water forms water pools which provide a breeding ground for mosquitoes, hence increasing the mosquito density which in turn leads to increase in malaria incidences. To our knowledge, no study has investigated the association between climate variables and malaria incidences in Apac District, Uganda.

The study had it's own limitations such as short data length and not being able to include non-climatic variables such as differences between human hosts, human migration and development projects which affect malaria transmission in the models. The relationship between malaria incidences and climate variables a period of 10 years was not found to be sufficient enough to predict future occurrences. Malaria incidence

is associated with socio-economic conditions of the people as well as malaria control measures. These factors were not incorporated in the models.

4.3. Results and Discussion on Forecasting of Malaria Incidences

4.3.1. Data Analysis

Monthly malaria incidences for the period $2007 - 2016$ was used. The data was obtained from the Ministry of Health, Uganda.

From Figures 4.6, 4.7 and 4.8, monthly malaria incidences, monthly rainfall and monthly average temperature from 2007 to 2016 were explored, which showed no clear trend and suggested a seasonal dependency in the series. All series exhibited number of peaks a part from small scale fluctuations. From Figure 4.6, the significant peaks in the monthly malaria incidences series seem to be separated by months showing a cyclical seasonal pattern as the peak of malaria incidences follow a similar pattern with an interval of few months between the peaks.

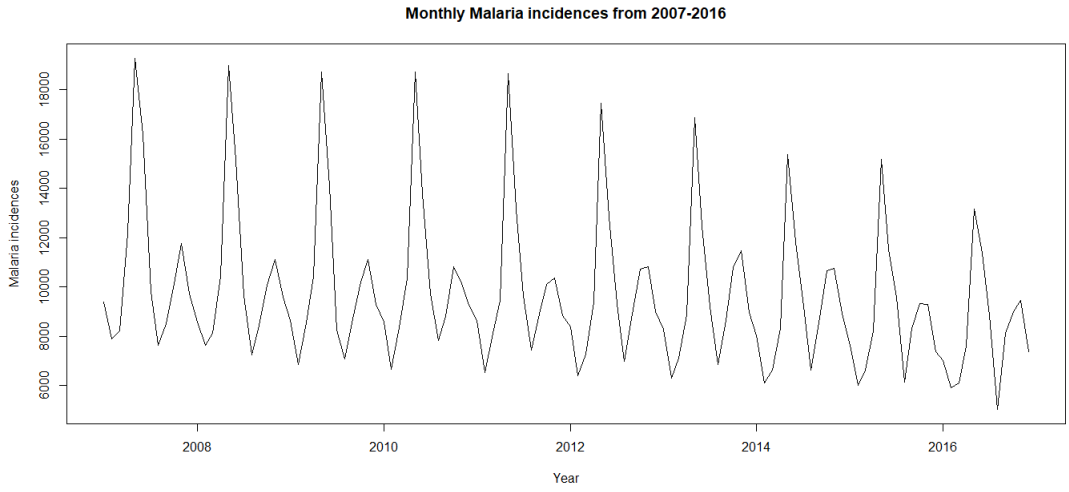

Figure 4.6.: Monthly Malaria incidences

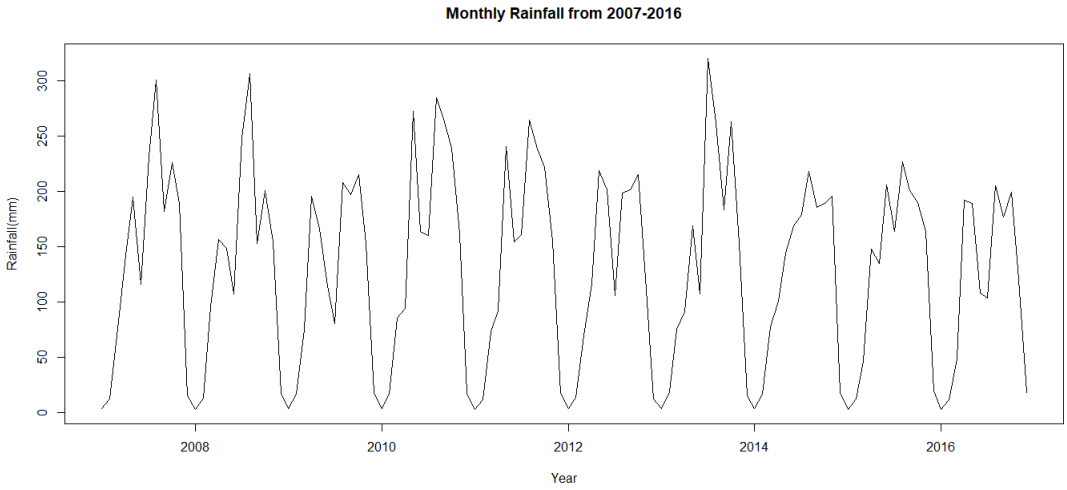

Figure 4.7.: Monthly Rainfall

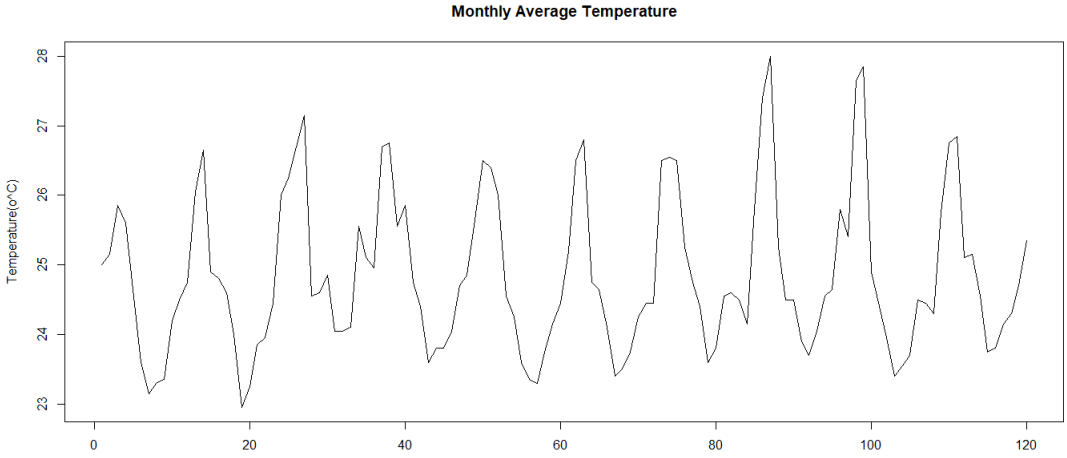

Figure 4.8.: Monthly Average Temperature

4.3.2. Results on Forecasting

The autocorrelation function(ACF) for the monthly malaria incidences series, Figure 4.9 showed significant peaks at different lags (autocorrelation=0.05918), Box-Ljung statistics (p-value=0.0000555695) as seen in Table 4.9, showed existence of serial correlation in the residuals of the model since the p-value is less than the 5% significance level. This indicates the mean of the residuals of the model were not finite and ARCH- LM test showed presence of arch effects(conditional Heteroscedasticity)

(p-value=0.0000413936) since the ARCH-LM test rejects the null hypothesis of no ARCH effect in the residuals of the model as the p-value of the Chi-square statistic is less than the 5% significance level as shown in Table 4.9. This indicates that the residuals are correlated thus do not follow white noise series. The series exhibited stationary (p-value=0.01) from the ADF test as observed in Table 4.9. The partial autocorrelation function (PACF), Figure 4.10 also showed significant peaks at different lags which confirmed presence of seasonal component in the time series data. From results in Table 4.9, it was noted that the actual data for malaria incidences exhibited presence of autocorrelation and arch effects in the residuals as well as lack of normality. This implied that the actual data malaria incidences series did not exhibit good behavior for forecasting purposes.

The data was transformed by performing a log transformation on the actual malaria incidences series. It was observed that the transformed data exhibited good behavior for forecasting purposes. The Ljung-Box test as seen in Table 4.10 showed that the residuals of the model were free from serial correlation since the p-values exceeds the 5% significance level. This indicates the mean of the residuals of the model were finite. Further ARCH-LM test shown in Table 4.10 showed that, the residuals of the model were free from conditional Heteroscedasticity, since the ARCH-LM test fails to reject the null hypothesis of no ARCH effect in the residuals of the model as the p-values of the Chi-square statistic is greater than the 5% significance level. This shows that the residuals are uncorrelated , thus have zero mean and have constant variance over time hence are white noise series.The residuals still showed lack of normality based on the p-values in Table 4.10. The transformed time series data showed was found to be stationary based on the results of the ADF test (p-value=0.01) in Table 4.10.

The R statistical software was used to find the best fit model for forecasting malaria incidences using monthly rainfall and monthly average temperature. It suggested the autoregressive integrated moving average model, ARIMA $(1,0,0)(1,1,0)^{12}$ as the best fit statistical model for this time series data. This is confirmed from results in Table 4.8. The observed values and the predicted values matched reasonably well as seen in Figure 4.12. The Ljung-Box statistics indicated that the model was specified correctly

as observed in Table 4.10.

The forecasting model proposed, ARIMA, provides a comprehensive set of tools for univariate time series model identification, parameter estimation and forecasting. It also offers great flexibility in analysis, which has contributed to it's popularity in several areas of research and practice. A seasonal ARIMA model is represented by ARIMA $(p, d, q)(P, D, Q)^s$ where p and P- are the autoregressive and seasonal autoregressive respectively, d and D- are the non-seasonal differences and seasonal differencing respectively, q and Q- are the moving average parameters and seasonal moving average parameters respectively, and s represents the length of the seasonal period.

ARIMA $(1, 0, 0)(1, 1, 0)^{12}$ model was used to forecast the malaria incidences for the future from January 2017 to December 2020. The fore casted malaria incidences also showed a seasonal pattern with significant peaks during the rainy season as observed in Figure 4.11.

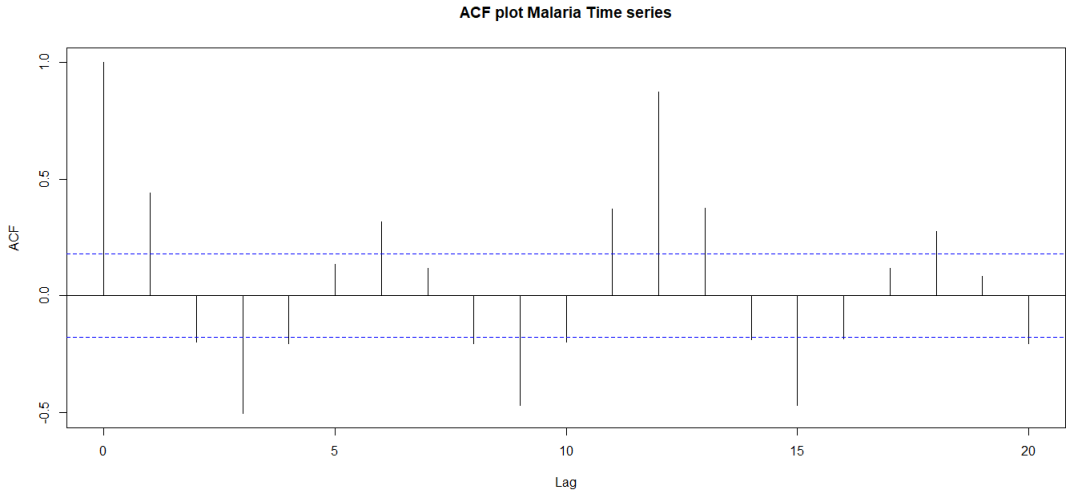

Figure 4.9.: ACF plot for Malaria incidences

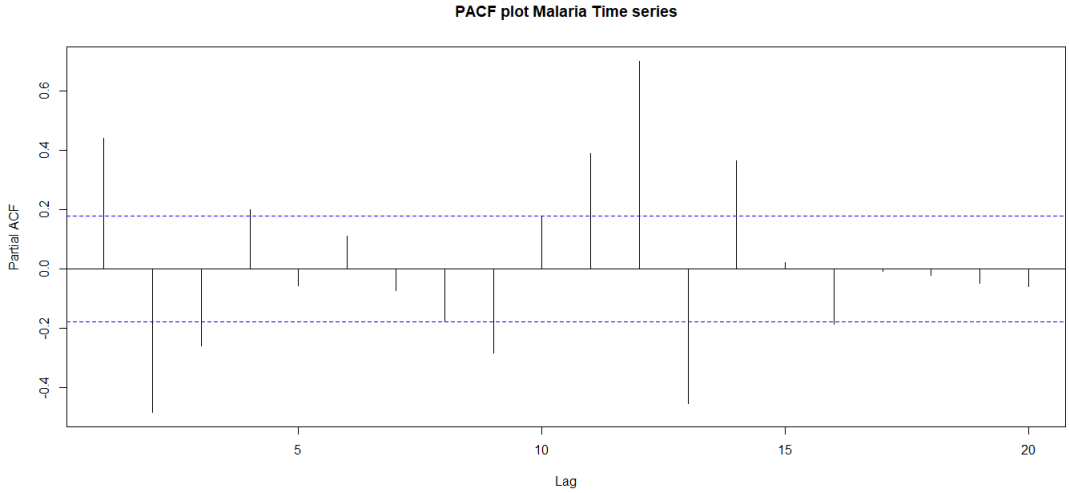

Figure 4.10.: PACF plot for Malaria incidences

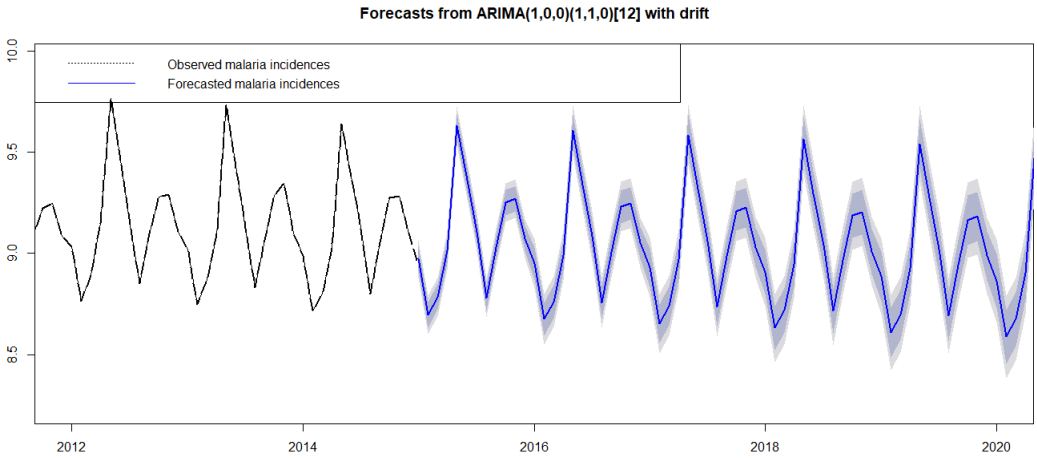

Figure 4.11.: Observed and Forecasted values for malaria incidences

Table 4.8.: Model statistics for malaria incidences

Model parameter	model type	AIC	BIC	MAE	ACF
Actual data	ARIMA$(0,0,3)$	1765.41	1778.23	1681.25	0.0591
Transformed data	ARIMA$(1,0,0)(1,1,0)[12]$	-268.85	-259.13	0.0307	-0.0301

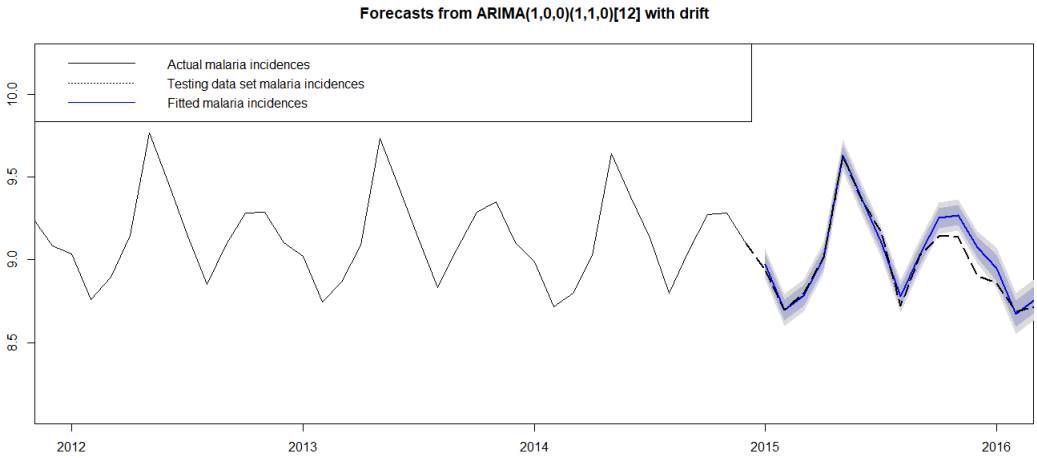

Figure 4.12.: Observed and fitted values for malaria incidences

Table 4.9.: Residual Analysis results for malaria incidences

Tests	Statistics	df	P-value
ARCH-LM test	95.913	20	$6.737e^{-12}$
Ljung-Box statistic	88.712	20	$1.244e^{-10}$
Jarque Bera Test	15.433	2	0.0004454
Augmented Dicker Fuller Test	-6.5335	lag$= 4$	0.01

Table 4.10.: Residual Analysis results for transformed malaria incidences

Tests	Statistics	df	P-value
Arch effects	17.54	20	0.6177
Ljung-Box statistic	29.18	20	0.08428
Jarque Bera Test	22.679	2	$1.189e^{-05}$
Augmented Dicker Fuller Test	-6.5352	lag$= 4$	0.01

4.3.3. Discussion on Forecasting Malaria incidences using ARIMA models

ARIMA models are useful tools in forecasting epidemiological data. They are particularly useful for diseases which show a seasonal pattern (Helfenstein, 1991), just like

in this study malaria incidences showed a seasonal pattern. ARIMA$(1,0,0)(1,1,0)^{12}$ model developed in this study is to provide a simple tool to predict the expected number of malaria incidences per month in the future based on observed malaria incidences over the years.

Wangdi et al. (2010), found ARIMA$(2,1,1)(0,1,1)^{12}$ to be the best possible model to predict malaria cases in Bhutan. The method of ARIMAX modelling was employed to determine predictors of malaria of the subsequent month. ARIMA model was also used for forecasting malaria cases in Sri Lanka (Briet et al., 2008) and Ethopia (Abeku et al., 2002).

High malaria incidences was found to occur between months of August, September and November. Rainy season in Apac District is usually from June, July and August. This shows that there is a strong correlation with rainfall in the preceding month. Rainfall increases the number of vector breeding grounds which is conducive to malaria transmission. Temperature also plays a significant role in malaria transmission. Temperature rise is expected to increase transmission and prevalence of malaria by shortening the incubation period of the parasite in the female Anopheles mosquitoes. Sporogonic cycles take about 9 to 10 days at temperatures of 28° but higher than 30° and below 16° have negative impact on parasite development (Hunter, 2003).

Based on the results of present study, we observe that malaria incidences will continue to occur in the near future based on forecasts made if appropriate actions are not initiated on time. The aim of this study was to develop a good forecasting model for predicting expected malaria incidences so that timely and control measures are put in place.

Apart from climatic factors, other factors like urbanization, population movement, the level of immunity to malaria in human hosts, insecticide resistance in mosquitoes and drug resistance in parasites play a significant role in affecting the malaria incidences. The statistical model developed in this study assumes these factors remain constant over a period of time taking into consideration only climatic factors to forecast malaria incidences.

The study had some limitations, the data used for the study was obtained from only

one health center. It makes it difficult to generalize results for the actual population due to small sample size. Non-climatic factors affecting malaria incidences were not included in the model. There was a challenge of obtaining weekly data; hence use of monthly data which affects the accuracy of the results.

5. Conclusions and Recommendations

5.1. Significance of climate variables on malaria incidences

Malaria remains an important public health problem in Apac District, Northern Uganda. The objective of this study was to model the climate variables that is rainfall and temperature associated with malaria incidences in Apac District. The study used monthly data for the period January 2007 to December 2016 in Apac District. The Poisson regression did not accurately fit the data on malaria incidences due to over dispersion in the data. The Negative Binomial Model was a better fit. The result obtained suggested that rainfall was positively significant on monthly malaria incidences whereas average temperature was not a significant predictor for malaria incidences based on results from Pearson correlation test in Apac District. A positive relationship between rainfall and expected malaria incidences was observed based on the coefficient value of parameter estimates in Table 4.5. The findings provide better insight of climate effects on malaria and provide important information for malaria prediction. It is observed that rainfall is a strong predictor of malaria incidences in Apac District. We recommend that in future studies, relative humidity, drug resistance, insecticide resistance in mosquitoes and control measures like use of treated insecticide mosquito nets should be incorporated in the models and more lengthy data set should be used.

5.2. Forecasting

A seasonal pattern was observed in the malaria incidences in Apac district. ARIMA $(1,0,0)(1,1,0)^{12}$ model was found to be the best fit statistical model to predict malaria incidences in Apac District. Rainfall was found to be a strong predictor of malaria incidences. The results found from this study offer useful information for policy makers to be able to effectively implement timely and effective malaria preventive and control measures.

We recommend that further research can be done by using data collected from more health centers and also non-climatic factors such as human migration, malaria control measures and land use be included in the model. We also recommend to evaluate the effectiveness of integrating the forecasting model into existing malaria control programme in terms of it's impact in reducing the disease occurrence and also the cost of control interventions.

References

Abeku, T. A., De Vlas, S. J., Borsboom, G., Teklehaimanot, A., Kebede, A., & Olana. (2002). Forecasting malaria incidence from historical morbidity patterns in epidemic-prone areas of Ethiopia: A simple seasonal adjustment method performs best. *Tropical medicine & international health*, *7*(10), 851–857.

Agresti, A. (2002). Inference for contingency tables. *Categorical Data Analysis, Second Edition*, *7*(1), 70–114.

Agusto, F. B., Marcus, N., & Okosun, K. O. (2012). Application of optimal control to the epidemiology of malaria. *Electronic Journal of Differential Equations*, *2012*(81), 1-22.

Akram, S., & Ann, Q. ul. (2015). Newton raphson method. *International Journal of Scientific & Engineering Research*, *6*(7), 2229-2241.

Al-Zeaud, H. A. (2011). Modelling and forecasting volatility using arima model. *European Journal of Economics, Finance and Administrative Sciences*, *35*(1), 109–125.

Aribodor, D., Ugwuanyi, I., & Aribodor, O. (2016). Challenges to achieving malaria elimination in Nigeria. *American Journal of Public Health Research*, *4*(1), 38-41.

Bloland, P., Boriga, D., Ruebush, T., McCormick, J., Roberts, J., Oloo, A., et al. (1999). Longitudinal cohort study of the epidemiology of malaria infections in an area of intense malaria transmission ii. descriptive epidemiology of malaria infection and disease among children. *The American journal of tropical medicine and hygiene*, *60*(4), 641-648.

Briet, O. J., Vounatsou, P., Gunawardena, D. M., Galappaththy, G. N., & Amerasinghe, P. H. (2008). Models for short term malaria prediction in Sri Lanka. *Malaria Journal*, *7*(1), 76-100.

Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data* (Vol. 53). Cambridge university press.

Chanda, E., Doggale, C., Pasquale, H., Azairwe, R., Baba, S., & Mnzava, A. (2013). Addressing malaria vector control challenges in South Sudan: proposed recommendations. *Malaria journal*, *12*(1), 59-70.

Connor, S., Thomson, M., & Molyneux, D. (1999). Forecasting and prevention of epidemic malaria: new perspectives on an old problem. *Parassitologia*, *41*(1-3), 439-448.

Contreras, J., Espinola, R., Nogales, F. J., & Conejo, A. J. (2003). Arima models to predict next-day electricity prices. *IEEE transactions on power systems*, *18*(3), 1014–1020.

Datta, K. (2011). Arima forecasting of inflation in the Bangladesh Economy. *IUP Journal of Bank Management*, *10*(4), 7-20.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, *74*(366), 427–431.

Dræbel, T., Kueil, B. G., & Meyrowitsch, D. W. (2013). Prevalence of malaria and use of malaria risk reduction measures among resettled pregnant women in South Sudan. *International health*, *5*(3), 211–216.

Dunn, P. K., & Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, *5*(3), 236–244.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the Econometric Society*, *50*(4), 987-1007.

Fahrmeir, L., & Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, *13*(1), 342–368.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *222*(2), 309–368.

Gomez-Elipe, A., Otero, A., Van Herp, M., & Aguirre-Jaime, A. (2007). Forecasting malaria incidence based on monthly case reports and environmental factors in Karuzi, Burundi, 1997-2003. *Malaria Journal*, *6*(1), 129-140.

Gujarati, D. N. (2009). *Basic econometrics*. Tata McGraw-Hill Education.

Health, M. of. (2014). *Uganda malaria reduction strategic plan 2014-2020*. Ministry of Health, Uganda.

Helfenstein, U. (1991). The use of transfer function models, intervention analysis and related time series methods in epidemiology. *International journal of epidemiology*, *20*(3), 808–815.

Hove-Musekwa, S. D., et al. (2008). Determining effective spraying periods to control malaria via indoor residual spraying in Sub-Saharan Africa. *Advances in Decision Sciences*, *2008*(1), 19-31.

Huang, F., Zhou, S., Zhang, S., Zhang, H., & Li, W. (2011). Meteorological factors–based spatio-temporal mapping and predicting malaria in Central China. *The American journal of tropical medicine and hygiene*, *85*(3), 560–567.

Hulden, L., & Hulden, L. (2009). The decline of malaria in Finland-the impact of the vector and social variables. *Malaria journal*, *8*(1), 94-106.

Hunter, P. (2003). Climate change and waterborne and vector-borne disease. *Journal of applied microbiology*, *94*(s1), 37-46.

Jarque, C. M., & Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review or Revue Internationale de Statistique*, *55*(2), 163-172.

Kakchapati, S., & Ardkaew, J. (2011). Modeling of malaria incidence in Nepal. *Journal of research in health sciences*, *11*(1), 7–13.

Kihoro, J., Otieno, R., & Wafula, C. (2004). Seasonal time series forecasting: A comparative study of ARIMA and ANN models. *African Journal of Science and Technology (AJST)*, *5*(2), 41-49.

Kim, Y.-M., Park, J.-W., & Cheong, H.-K. (2012). Estimated effect of climatic variables on the transmission of plasmodium vivax malaria in the Republic of Korea. *Environmental health perspectives*, *120*(9), 1314-1326.

Krefis, A. C., Schwarz, N. G., Krüger, A., Fobil, J., & Nkrumah. (2011). Modeling the relationship between precipitation and malaria incidence in children from a holoendemic area in Ghana. *The American journal of tropical medicine and hygiene*, *84*(2), 285–291.

Kumar, K., Yadav, A., Singh, M., Hassan, H., & Jain, V. (2004). Forecasting daily maximum surface ozone concentrations in Brunei Darussalaman arima modeling approach. *Journal of the Air & Waste Management Association*, *54*(7), 809–814.

Kumar, V., Mangal, A., Panesar, S., Yadav, G., Talwar, R., Raut, D., et al. (2014). Forecasting malaria cases using climatic factors in Delhi, India: a time series analysis. *Malaria research and treatment*, *2014*(1), 6-18.

Lawless, J. F. (1987). Negative binomial and mixed poisson regression. *Canadian Journal of Statistics*, *15*(3), 209–225.

Lindsay, S. W., Bødker, R., Malima, R., Msangeni, H. A., & Kisinza, W. (2000). Effect of 1997–98 ei niño on highland malaria in Tanzania. *The Lancet*, *355*(9208), 989–990.

Liu, Q., Liu, X., Jiang, B., & Yang, W. (2011). Forecasting incidence of hemorrhagic fever with renal syndrome in China using arima model. *BMC infectious diseases*, *11*(1), 218-227.

Ljung, G. M., & Box, G. E. (1978). On a measure of lack of fit in time series models. *Biometrika*, *65*(2), 297-303.

Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (2008). *Forecasting methods and applications, Third Edition*. John wiley & sons.

Martens, P., & Hall, L. (2000). Malaria on the move: human population movement and malaria transmission. *Emerging infectious diseases*, *6*(2), 103-115.

McCullagh, P., & Nelder, A., J. (1992). *Generalized linear models. in breakthrough in statistics*. Springer New York.

McLeod, A. I., & Li, W. K. (1983). Diagnostic checking arma time series models using squared-residual autocorrelations. *Journal of Time Series Analysis*, *4*(4), 269–273.

McMichael, A. J., Woodruff, R. E., & Hales, S. (2006). Climate change and human health: present and future risks. *The Lancet*, *367*(9513), 859–869.

Millennium, U. N. U. (2005). Coming to grips with malaria in the new millennium, task force on hiv/aids, malaria, tb and access to essential medicines. *Journal of the United Nations*.

Muggeo, V. M. (2008). Modeling temperature effects on mortality: multiple segmented relationships with common break points. *Biostatistics*, *9*(4), 613–620.

Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, *24*(3), 69–71.

Musa, M. I., Shohaimi, S., Hashim, N. R., & Krishnarajah, I. (2012). Environmental and socio-economic determinants of malaria rate in Sudan. *Research Journal of Environmental and Earth Sciences*, *4*(11), 923–929.

Nath, D. C., & Mwchahary, D. D. (2013). Association between climatic variables and malaria incidence: A study in Kokrajhar District of Assam, India: Climatic variables and malaria incidence in Kokrajhar District. *Global journal of health science*, *5*(1), 90-98.

Nkurunziza, H., Gebhardt, A., & Pilz, J. (2010). Bayesian modelling of the effect of climate on malaria in Burundi. *Malaria Journal*, *9*(1), 114-125.

Nkurunziza, H., Gebhardt, A., & Pilz, J. (2011). Geo-additive modelling of malaria in Burundi. *Malaria journal*, *10*(1), 234-246.

Okello, V. B., Byaruhanga, C., Roelants, T., & DAlessandro, C. (2006). Variation in malaria transmission intensity in seven sites throughout Uganda. *The American journal of tropical medicine and hygiene*, *75*(2), 219–225.

Park, H. (1999). *Forecasting three-month treasury bills using ARIMA and GARCH Models.* Econometrics.

Pascual, M., Ahumada, J. A., Chaves, L. F., Rodo, X., & Bouma, M. (2006). Malaria resurgence in the East African highlands: temperature trends revisited. *Proceedings of the National Academy of Sciences*, *103*(15), 5829-5834.

Patience, E. O., & Osagie, A. M. (2014). Modeling the prevalence of malaria in Niger State: An application of poisson regression and negative binomial regression models. *International Journal of Physical Sciences*, *4*(2), 061-068.

Paul, J. C., Hoque, M. S., & Rahman, M. M. (2013). Selection of best arima model for forecasting average daily share price index of pharmaceutical companies in Bangladesh: A case study on square pharmaceutical ltd. *Global Journal of Management And Business Research*, *13*(3), 1–13.

Rogers, D. J., & Randolph, S. E. (2000). The global spread of malaria in a future, warmer world. *Science*, *289*(5485), 1763-1766.

Sachs, J., & Malaney, P. (2002). The economic and social burden of malaria. *Nature*, *415*(6872), 680-685.

Shanks, G. D., Hay, S. I., Stern, D. I., Biomndo, K., & Snow, R. W. (2002). Meteorologic influences on plasmodium falciparum malaria in the highland tea estates of Kericho, western Kenya. *Emerging infectious diseases*, *8*(12), 1404-1408.

Sriwattanapongse, W., Me-ead, S., & Khanabsakdi, S. (2011). Forecasting malaria incidence based on monthly case reports and climatic factors in Ubon Ratchathani Province, Thailand, 2000-2009. *Science Journal Ubon Ratchathani University*, *2*(1), 17-30.

Talisuna, A. O., Okello, P. E., Erhart, A., Coosemans, M., & DAlessandro, U. (2007). Intensity of malaria transmission and the spread of plasmodium falciparum–resistant malaria: a review of epidemiologic field evidence. *The American journal of tropical medicine and hygiene*, *77*(6), 170-180.

Teklehaimanot, H. D., Lipsitch, M., Teklehaimanot, A., & Schwartz, J. (2004). Weather-based prediction of plasmodium falciparum malaria in epidemic-prone regions of Ethiopia I. patterns of lagged weather effects reflect biological mechanisms. *Malaria journal*, *3*(1), 41-52.

Tsitsika, E. V., Maravelias, C. D., & Haralabous, J. (2007). Modeling and forecasting pelagic fish production using univariate and multivariate arima models. *Fisheries science*, *73*(5), 979–988.

Uko, A. K., & Nkoro, E. (2012). Inflation forecasts with arima, vector autoregressive and error correction models in Nigeria. *European Journal of Economics, Finance & Administrative Science*, *50*(1), 71–87.

Vatcheva, K. P., Lee, M., McCormick, J. B., & Rahbar, M. H. (2016). Multicollinearity in regression analyses conducted in epidemiologic studies. *Epidemiology (Sunnyvale, Calif.)*, *6*(2), 51-65.

Venables, W. N., & Ripley, B. D. (2002). Modern applied statistics. Springer.

Wangdi, K., Singhasivanon, P., Silawan, T., Lawpoolsri, S., White, N. J., & Kaewkungwal, J. (2010). Development of temporal modelling for forecasting and

prediction of malaria infections using time-series and arimax analyses: a case study in endemic Districts of Bhutan. *Malaria Journal*, *9*(1), 251-260.

Wardrop, N. A., Barnett, A. G., Atkinson, J.-A., & Clements, A. C. (2013). Plasmodium vivax malaria incidence over time and its association with temperature and rainfall in four counties of Yunnan Province, China. *Malaria journal*, *12*(1), 452-465.

WHO. (2004). Report of an informal consultation, world health organization [Computer software manual]. World Health Organization.

WHO. (2013). *Who report*. World Health Organization Report.

WHO. (2014). World health day 2014: Preventing vector-borne disease. *http://www.who.int/campaigns/world-health-day/2014/en/*.

WHO. (2015). *Who report*. World Health Organization Report.

WHO. (2017). World health organization(who) and who global malaria programme [Computer software manual]. World Health Organization. Available from `http://www.who.int/mediacentre/factsheets/fs094/en/`

Williams, R. (2016). *Models for count outcomes*. University of Notre Dame, USA.

Yang, H. M., & Ferreira, M. U. (2000). Assessing the effects of global warming and local social and economic conditions on the malaria transmission. *Revista de saude publica*, *34*(3), 214-222.

Yeka, A., Gasasira, A., Mpimbaza, A., Achan, J., Nankabirwa, J., & Nsobya. (2012). Malaria in Uganda: challenges to control on the long road to elimination: I. epidemiology and current control efforts. *Acta tropica*, *121*(3), 184–195.

Zhou, G., Minakawa, N., Githeko, A. K., & Yan, G. (2004). Association between climate variability and malaria epidemics in the East African highlands. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(8), 2375-2380.

Appendices

Appendix

A. Properties of the Poisson random variable

Theorem A.1. *The probability mass function;* $f(x) = \frac{e^{-\lambda}\lambda^x}{x!‘}$ *for a Poisson random variable* X *is a valid p.m.f.*

Proof. $f(x) = \frac{e^{-\lambda}\lambda^x}{x!‘}$ for $x = 0, 1, 2, ...$

1. $f(x) > 0$ because $\lambda^x > 0, e^{-\lambda} > 0, x!‘ > 0$
2. $\sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!‘} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!‘} = e^{-\lambda}[1 + \frac{\lambda}{1!‘} + \frac{\lambda^2}{2!‘} + ...] = e^{-\lambda}e^{\lambda} = 1$

□

Theorem A.2. *The moment generating function of a Poisson random variable* X *is:* $M(t) = e^{\lambda(e^t-1)}$ *for* $-\infty < t < \infty$.

Proof. $M(t) = E(e^{tX})$ by definition
$= \sum_{x=0}^{\infty} e^{tx}.\frac{e^{-\lambda}\lambda^x}{x!‘} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!‘} = e^{-\lambda}.e^{\lambda e^t} = e^{\lambda(e^t-1)}$ □

Theorem A.3. *The mean of a Poisson random variable* X *is* λ.

Proof. $M(t) = e^{\lambda(e^t-1)} = M'(t) = e^{\lambda(e^t-1)}.\lambda e^t = M'(0) = e^{\lambda(e^0-1)}.\lambda e^0 = e^{\lambda(1-1)}.\lambda(1) = \lambda$ □

Theorem A.4. *The variance of a Poisson random variable* X *is* λ.

Proof. $M'(t) = e^{\lambda(e^t-1)}.\lambda e^t = M''(t) = e^{\lambda(e^t-1)}.\lambda e^t + M'(t) = \lambda e^t.e^{\lambda(e^t-1)}.\lambda e^t = M''(0) = e^{\lambda(e^0-1)}.\lambda e^0 + \lambda e^0.e^{\lambda(e^0-1)}.\lambda e^0 = \lambda + \lambda^2$
$\sigma^2 = M''(0) - (M'(0))^2 = \lambda + \lambda^2 - \lambda^2 = \lambda$ □

B. R codes

```
> dat<-read.csv(file.choose(),header=TRUE)\\
> library(MASS)\\
> fit1<-glm.nb(dat$Mal~dat$rain+AvgTemp , data=dat)\\
> summary(fit1)\\
> fit1<-glm.nb(dat$Mal~AvgTemp , data=dat)\\
> summary(fit1)  \\
> qqnorm(fit1$residuals)\\
 $>$ qqline($fit1$residuals)\\
> fit1<-glm.nb(dat$Mal~dat$rain , data=dat)
> summary(fit1)
> qqnorm(fit1$residuals)
> qqline(fit1$residuals)
> cor(dat$Mal,dat$rain)
> cor(dat$rain,dat$AvgTemp)
> hist(dat$Mal,main="Monthly Malaria incidences", xlab="", ylab="Ma
> fit1<-glm(dat$Mal~dat$rain+dat$AvgTemp, family=poisson(link="log"
> summary(fit1)
> qqnorm(fit1$residuals)
> qqline(fit1$residuals)
> plot(dat$Mal,main="Monthly malaria incidences", xlab="",ylab="Numb
> acf(dat$Mal,main="ACF plot Malaria Time series")
> pacf(dat$Mal,main="PACF plot Malaria Time series")
> auto.arima(dat$Mal)
> data.test=window(ts.data,start=c(2015,1))
> data.train=window(ts.data,start=c(2007,1),end=c(2014,12))
> ts.data=ts((dat$Mal),frequency=12,start=c(2007,1))
> plot(ts.data)
> library(TSPred)
> plot(arima.forecast,xlab="years",ylab="malaria incidences")
```

```
> plotarimapred(data.test,arima1,xlim=c(2012,2018),range.percent=0.
> plotarimapred(data.test,fit,xlim=c(2012,2016),range.percent=0.05)
> lines(predfit$pred,col="blue")
> library(tseries)
> jarque.bera.test(arima2$residuals)
> adf.test(log(dat$Mal))
> fit=auto.arima(dat$Mal)
> arima2=auto.arima(data.train)
> summary(arima2)
> Box.test(arima2$residuals^2,lag=20,type="Ljung-Box")
> Box.test(arima2$residuals,lag=20,type="Ljung-Box")
> plot.ts(arima2$residuals)
> acf(arima2$residuals,lag.max=24,main="ACF")
> fit1=auto.arima(log(dat$Mal))
> plot.ts(fit$residuals)
> Box.test(fit$residuals^2,lag=20,type="Ljung-Box")
> acf(fit1$residuals,lag.max=24,main="ACF")
> Box.test(fit1$residuals,lag=20,type="Ljung-Box")
> plotarimapred(data.test,arima2,xlim=c(2012,2018),range.percent=0.
> ts.data=ts(log(dat$Mal),frequency=12,start=c(2007,1))
> plot(ts.data)
> data.train=window(ts.data,start=c(2007,1),end=c(2014,12))
> data.test=window(ts.data,start=c(2015,1))
> library(forecast)
> arima1=auto.arima(data.train)
> summary(arima1)
> plot.ts(arima1$residuals)
> arima.forecast=forecast(arima1,h=24)
> plot(arima.forecast,xlab="years",ylab="malaria incidences")
> library(TSPred)
```

```
> plotarimapred(data.test,arima1,xlim=c(2012,2018),range.percent=0.
```