

MOI UNIVERSITY

PHD THESIS

Flexible Models for Analyzing Correlated
and Non-Normal Data with Application to
Health Research

DPS/PHD/009/2017

Ngugi MWENDA

*A thesis submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Biostatistics*

in the

School of Science and Aerospace Studies


Department of Mathematics, Physics and Computing

November 2021



Declaration

This thesis is my original work and has not been submitted for a degree in any other institution. No part of this thesis may be reproduced without prior permission of the author and/or Moi university

Signed:  Date: 29-11-2021

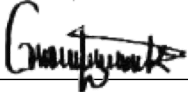
Ngugi Mwenda

Supervisors

Signed:  Date: 29-Nov-2021

Dr Mathew Kosgei,

Department of Mathematics, Physics and Computing, Moi University

Signed:  Date: 29/11/2021

Dr Gregory Kerich,

Department of Mathematics, Physics and Computing, Moi University

Signed:  Date: 29-Nov-2021

Prof. Ruth Nduati,

Department of Pediatrics, University of Nairobi

Dedication

To the memory of my Auntie Selina Kagendo

To my main guy Jawara Samwel and the love of my grandparents

Acknowledgements

First of all, I would want to thank my three supervisors, Dr Mathew Kosgei, Dr Gregory Kerich and Prof Ruth Nduati for their dedication and time to review this work.

I would wish to thank former ICT director at KNBS Mr. Cleophas Kiiro, who advised me to go for a PhD when I was thinking of enrolling for a second masters. Your advice has paid off finally. Much thanks to former Director General at KNBS Mr Zachary Mwangi, who ensured a smooth environment for the study. I would also wish to appreciate the Director General Mr Macdonald Obudho for always believing in me. I also wish to thank my immediate bosses Mr Silas Mulwa, Mr James Nganga and Mr John Bore and also senior manager for research Mr. Paul Samoei and finally the director incharge Mr. Benjamin Avusevwa for giving a conducive environment for study.

Special thanks to Mr Kariuki, Principal Iruma Mixed day Secondary and former Deputy Principal Chogoria Boys 2005. Mwalimu you ensured I got an education against all odds on lack of school fees, may God increase you and always bless your kind heart.

Thanks to Ev.Paul Kirimi for your continued support in prayer and guidance. Thanks for standing with me whenever I called.

To Pst. Kato, you always find a special place in my heart. You have been a great friend and encourager.

To Mr Mutua Kakinyi, much appreciated for providing the data sets. Mr Buluma Robert you have been a great asset to me and a big encouragement. Thanks Joseph Kombe of Taita University for running the R codes and Noah Mutai for proof reading my papers.

In absentia, much thanks to Prof. Taryn Swan for your general contribution to open source code knowledge sharing to fit the models. I thank Mr SMT Wambua for being that person who corrected part of my thesis.

Dr C. K. Laboso of ILRI, you have constantly been a big cheer leader whenever i shared the progress of the report with you. You are an amazing inspiration.

Thankyou soo much Dr P. N. Kariuki for the good wishes, the endless discussion on how to become a better doctor cannot go unmentioned. To Banker P. W. Wambugu, the friendship we started and the kind words you have always spoken have motivated me to be a better person.

To all my comrades of 2012 at Kenya National Bureau of Statistics, whom we started a lasting friendship to date. Your words of encouragement have a lasting mark in my life. Finally, I thank all my Ph.D classmates at Moi university for always being greatest motivators.

Special thanks to all my family members, particularly my grandfather M'amwenda Chabari and grandmother Elizabeth Mwenda, you ensured I get education. May you live long. To my mum Jane and dad Allan, thankyou for being the best parents ever.

We acknowledge the Forgy International Center, National Institute of Health, which funded the first clinical trial through grant registration number D43-TW00007 and T22-TW00001.

Contents

Dedication	ii
Acknowledgements	iii
Contents	v
List of Figures	viii
List of Tables	ix
Abbreviations	xi
Symbols	xiii
Abstract	xiv
1 Introduction	1
1.1 Background	1
1.2 Statement of the Problem	5
1.3 Justification	7
1.4 Objectives	8
1.5 Organization of the Thesis	8
2 Literature Review	9
2.1 Introduction to Literature Review	9
2.2 Bacterial Vaginosis (BV)-HIV-1 Co-Existence on Maternal and Infant health	9
2.3 Non-Normality and Correlation in an Infant Morbidity Longitudinal study	14
2.4 Outpatient Care Cost in Kenya	17
2.5 Distance Traveled for Inpatient Care in Kenya	20
3 Methodology	27
3.1 Introduction	27
3.2 Suppressor Effect Application for Modelling Correlation under Independence Assumption	27
3.2.1 Study population, enrollment, delivery, and follow up	28

3.2.2	Clinical characteristics	30
3.2.3	Laboratory methods for detection of Bacterial vaginosis	32
3.2.4	Ethical Approval	32
3.2.5	Statistical analysis	32
3.3	Skewed Logit model for correlated binomial longitudinal data and application to modelling infant morbidity under HIV setting	34
3.3.1	Materials and methods	34
3.3.2	Statistical Model	37
3.3.3	Estimation of parameters using the GEE	40
3.4	Tweedie Distribution for a Response Exhibiting is continuous, non-negative and Right Skewed Characteristic Using Independent Structure with application to health cost data	44
3.4.1	Exploratory Data Analysis	44
3.4.2	Tweedie Distribution	46
3.4.3	Mean Function	48
3.4.4	Variance function	48
3.4.5	Approximating tweedie densities using saddle point approximation	52
3.4.6	Estimating the Parameters	53
3.4.7	Working correlation	57
3.4.8	Independent Correlation Structure for Imbalanced Data	58
3.4.9	Models selection	59
3.5	Tweedie distributions in modeling clustered data using exchangeable correlation structure and applications to distance-for inpatient care data	61
3.5.1	Notations	62
3.5.2	Exchangeable or Symmetrical Correlation for imbalanced data	63
4	Results and Discussion	69
4.1	Introduction	69
4.2	Effect of Bacterial Vaginosis (BV)-HIV-1 Co-Existence on Maternal and Infant health: A Secondary Data Analysis Results	69
4.2.1	Mortality in the first twelve months of life	75
4.2.2	Discussion	75
4.3	Skewed Logit Model for Correlated data and its application to infant morbidity results	81
4.3.1	Effects of time on BV	85
4.3.2	Discussion	86
4.3.3	Model Diagnostics	90
4.4	Predictors for Outpatient Care Cost in Kenya	92
4.4.1	Model Results	92
4.4.2	Discussion	96
4.5	Results from Analysis of Distance Traveled for Inpatient Care in Kenya	106
4.5.1	Model selection	106
4.5.2	Discussion	108
4.5.3	Model Diagnostics	117
4.5.4	A non-parametric test of the randomness of residuals	117

4.5.5	Raw residuals analysis using Plots and graphical assessments . . .	119
5	Conclusions, Recommendations and Further Research	123
	Bibliography	129
	Appendix A: Obtaining the disturbance term α for use in section 4.3	143
	Appendix B: Proof for alternative calculation of scale parameter by altering the denominator under Exchangeable correlation	154
	Appendix C: R codes	157
	Appendix D: Supplementary Tables and Figures	190
	Appendix E: List of Publications and Conferences	193

List of Figures

3.1	<i>Trial flow diagram describing the recruitment of cases and exclusions . . .</i>	30
3.2	<i>Cumulative density function of the skewed logit model with different values of skewness. The bold continuous line represents the logit model which assumes symmetry.</i>	39
3.3	<i>The profile log-likelihood plot for cost of out patient care in Kenya using the model 1 covariates. The solid line is a saddle-point approximation of the P index from the data with a value of 1.68 and estimated 95% CI [1.67,1.69]</i>	61
3.4	<i>The profile log-likelihood plots for the distance travelled for inpatient care in Kenya using gender of the household head, household size, and education as covariates. The plot estimated the p as 1.63 (1.59,1.67), with the dots representing 95%. The solid line is a cubic-spline smooth interpolation joining the points.</i>	68
3.5	<i>Histogram of Distance travelled to seek inpatient care in Kenya</i>	68
4.1	<i>Kaplan–Meier analysis of infant mortality over 12 months between infants with maternal exposure/ non-exposure to bacterial vaginosis</i>	76
4.2	<i>Variations of mean cost for out-patient by household head age</i>	95
4.3	<i>Bivariate plot of mean cost for mean out-patient costs by household head age for regions in Kenya</i>	104
4.4	<i>Bivariate plot of median cost for out-patient costs by household head age for regions in Kenya</i>	105
4.5	<i>Plot of raw residuals versus the observation numbers</i>	120
4.6	<i>QQ normal Plots</i>	121
4.7	<i>Plot of raw residuals versus the Linear Predictor</i>	121
1	<i>Infant mortalities for one year between infants whose mothers are exposed to the BV and unexposed</i>	190
2	<i>Weight gain for age for the exposed and unexposed</i>	191

List of Tables

3.1	<i>Distributions of selected members of Exponential families</i>	47
3.2	<i>Quasi Likelihood Distributions for selected members of Exponential Families Densities</i>	49
3.3	<i>Descriptive analysis of Distance for inpatient care</i>	65
4.1	Comparison of the Mothers Demographic and selected maternal Characteristics between the two Groups	71
4.2	<i>Outcomes of maternal morbidity incidences</i>	72
4.3	<i>OR of bacterial vaginosis and Log viral load</i>	72
4.4	<i>Distribution of neonatal characteristics</i>	72
4.5	<i>Bacterial vaginosis and birth weight as a continuous variable</i>	73
4.6	<i>ORs of morbidity incidence among the neonates</i>	73
4.7	<i>Overall morbidity incidences reported and correlation analysis of bacterial vaginosis and morbidity incidences</i>	74
4.8	Predictors of bacterial vaginosis at 6 and 12 months with corresponding 95% Confidence Intervals (CI) and <i>p</i> -values	75
4.9	<i>BV with morbidity incidences reported from month one to six for both BV-exposed and unexposed babies in the Nairobi data survey</i>	82
4.10	<i>Regression Parameter Estimates with Model-Based and Empirical Standard Errors (SE) for Independence, Exchangeable, AR(1), Unstructured and M-dependent Correlation Structures Estimated Using Unconditional Residuals for GEE and skewed logit-GEE</i>	84
4.11	<i>Calculated coefficient of bacterial vaginosis with time from $\exp(\beta_1 + \beta_{15} \times \text{time})$, achieved by replacing the respective values from the skewed logit-GEE model with the AR-1 correlation structure</i>	85
4.12	<i>Differences in Model-based vs Sandwich-based Variance Ratios for both GEE and SL-GEE</i>	91
4.13	<i>A summary of the total cost for outpatient care incurred by the households from the KHHEUS 2018 data. Statistics have been recorded for the survey month total cost ≥ 0 KES (all the households) and survey month cost > 0 KES (Those who spend money on outpatient care only). All measurements are recorded in Kenya Shillings(KES).</i>	93
4.14	<i>Different model outputs with calculated QICu. The model with the lowest QICu is selected as the best fitting model. In our case, Model 1 is selected as the most parsimonious model for predicting outpatient care cost among households in Kenya using the Kenya Household Health Utilization Survey 2018</i>	93

4.15	<i>The residual deviance and degrees of freedom for a Tweedie glm with differing link functions using Model 1 covariates</i>	100
4.16	<i>Models selection using QICu and R^2</i>	107
4.17	<i>The residual deviance and degrees of freedom for a Tweedie glm with differing link functions using Model 7 covariates</i>	109
4.18	<i>Models Diagnostics using Wald–Wolfowitz run test</i>	118
1	<i>Run time in seconds in calculating the denominator of the scale parameter</i>	156
2	<i>Compute the skewness parameter(α)</i>	192

Abbreviations

BV	B acterial V aginitis
GEE's	G eneralized E stimating E quations
GLM	G eneralized L inear M odels
HIV	H uman I mmunodeficiency V irus
STD's	S exually T ransmitted D iseases
VL	V iral L oad
CD4	C luster of D ifferentiation 4
CD8	C luster of D ifferentiation 8
APGAR	A ppearance, P ulse, G rimace, A ctivity, and R espiration score
OR	O dd's R atio
s.e	standard error
R	R Statistical Software
MDG's	M illennium D evelopment G oals
AR-1	A utoregressive -1
GLMM	G eneralized L inear M ixed M odel
SL-GEE's	S kewed L ogit - G eneralized E stimating E quations
CL-GEE's	C onventional L ogit - G eneralized E stimating E quations

VCE	Variance C luster E stimate
IND	Independen t correlation
EDM	Exponential D ispersion M odels
EXCH	Exch e ngable correlation
M-DEP	M-de p endent correlation
UNSTR	Un s tructured correlation
MCTC	Mother to C hild T ransmission
IRLS	Iterative R e-Weighted L east S quare
QIC	Quasi I nformation C riteria
QICu	Quasi I nformation C riteria under Independence Assumptions
CI	Confidence I nterval
VR	Variance R atio
LBW	Low B irth W eight
MLE	Maximum L ikelihood E stimate
CDF	Cumulative D ensity F unctions
M-CAR	Missing C ompletely at R andom
NIMS	Nairobi I nfant M orbidity S tudy

Symbols

α	Significance Level	0.05
CI	Confidence Interval	95%
Y_i	$n \times 1$ Vector Matrix of responses	
X_i	$n \times p$ Vector Matrix of covariates	
β 's	coefficient of covariates	
∂	partial derivative	
φ	Disturbance term	
$k(.)$	Link function	

Abstract

Skewed and non-normal data are commonly observed in health research. Usually, the dataset is transformed, censored, or truncated to impose normality, rather than modeling the data in its natural state. Many conventional approaches to modeling lead to incorrect estimates of parameters and standard errors due to the assumptions imposed. This study investigated three statistical problems; Non-normality, Skewness and Correlation under the Generalized Estimating Equations (GEEs) Framework. The general objective was to develop flexible models for correlated and Skewed data in health research. The specific objectives were, to investigate the skewness property of binomial longitudinal data and its application to model infant morbidity under HIV setting, to review cost spending models on outpatient care while assuming independence structure under GEE, to develop models for alternative estimation of the scale parameter under the Generalized Estimating Equations framework, and to propose methods of analyzing clustered inpatient care data that relaxes the non-normality. The study applied; the Burrs-10 distribution and suppressor effect assumptions under the GEE to model the correlated infant morbidity data; exchangeable correlation structure to model predictors of distance for inpatient care; and independence correlation structure to model the predictors for outpatient care cost with Bienayme–Chebyshev inequality. The best model selected was the one that displayed the lowest quasi-likelihood under the independence criterion (QICu). The results revealed that skewed logit-GEE under the Burrs-10 distribution was able to show an association between variables which was not identified by the standard GEE. Accordingly, it fitted our imbalanced health dataset better. The study found out that the SL-GEE was superior over the standard GEE when asymmetry was assumed. The main contribution of the

study is in the development of the algorithm of estimating the skewness parameter for the model. The suppressor effect showed some patterns of the disease, which the conventional approaches failed to reveal. It revealed that gastro intestinal infections were common in the infants exposed to Bacterial Vaginosis. Modelling the distance for inpatient care revealed that differences in employment, ability to pay for the service and household size are associated with distance covered to access government facilities. Finally, the best predictors of outpatient care expenses are age, wealth index, marital status, and education of the household head. In conclusion, the methodologies developed are applicable in modelling of non-normal response variable. The study recommends the reproducibility of the R-codes developed on different health and biomedical datasets.

Chapter 1

Introduction

1.1 Background

Standard questionnaires are developed by researchers in an attempt to collect data that are meant to answer certain scientific questions in health research. This includes studies such as Demographic Health Surveys and the Household health expenditure surveys. Most of the outcomes of study are either continuous or dichotomous.

In a research on health status of a given population, a response of whether a person experienced any sickness within a given time period, say a year, can be captured as a yes or no [[KNBS, 2015](#)] and in the same research, distance covered by the patient or cost incurred at the facility to treat the illness can be captured as a continuous variable [[GOK, 2014](#)].

Some of the data collected, though may be of the said form, may not be fully modeled under the given conventional assumptions. Some binary responses may exhibit asymmetry, whereby the sensitivity to changes in the independent variable is not maximized at 0.5.

The GLM framework can be extended by including a skewness parameter to relax this imbalance [Nagler, 1994, Prentice, 1976]. The model relaxes the sensitivity to change in the dependent variable from 0.5 to a value that is determined by the data. In this case the stimuli in any of the independent variables for any individual with $p=0.5$ is not exaggerated.

Longitudinal data involves taking measures at different points from the same subject over time. This gives rise to data that exhibit correlation. This correlation should be factored in during analysis to gain meaningful results. Under the GLM framework, random effects models are good candidates for this type of analysis [Ibrahim et al., 2010, Rizopoulos et al., 2017].

However, there are several restrictive assumptions of the GLM, that make them undesirable. For instance, specifying the full likelihood can be ineffective if some information regarding it are unavailable. Methods that relax these assumptions have been developed such as the Generalized Estimating Equations (GEE) [Liang and Zeger, 1986].

Models that combine the aspect of skewed correlated responses are desirable. An extension to accommodate both have been developed by McDaniel et al. [2013]. This is through a two- stage analysis, whereby in first stage, the skewness parameter is specified while in the second stage, the value is applied under the GEE framework.

Hitherto, the research has discussed one angle of the response, that is the binary under a longitudinal study. Some data may be clustered and exhibit correlation among the subjects. When the response is continuous, the data can exhibit non-normality. For example, data on inpatient care for patients who traveled to seek care in government hospitals. Those who are within reach didn't cover any distance to the facility, some covered a moderate distance while others covered very long distances.

Evaluating this response in a plot clearly showed that the data; (1) have a discrete mass at zero for those within reach and (2) are right skewed for those who covered very long distances. Such data being collected from the same clusters, say counties, have respondents exhibiting similar characteristics thus correlated.

It is clear this response is complicated to model, and several attempts have been suggested in the literature. An extension of the Poisson to zero inflated Poisson [[Lambert, 1992](#)] has been found attractive, however, its ability to handle skewness has been put to doubt. Another suggestion to handle the continuous right skewed data is the Gamma model [[Agarwal and Kalla, 1996](#)].

An attempt to improve on this method was an extension by [[Dunn, 2017](#), [Dunn and Smyth, 2005, 2008](#), [Gilchrist and Drinkwater, 2000](#), [Smyth and Jørgensen, 2002](#)], in which under their method, they assumed the data had both Poisson-gamma characteristics. A Bayesian approach incorporating priors under the tweedie distribution was proposed by [Swallow et al. \[2016\]](#).

However, there have been few methodologies that have considered correlation and few researchers have attempted to improve on the methods based on GEE [[Liang and Zeger,](#)

1986].

The approach to model non-normal responses is gaining interest among researchers, with numerous methods considered of late [Bono et al., 2017], and this forms the basis of this thesis.

The focus on modelling correlated binary outcomes, proposed an improved algorithm for handling asymmetrical binary responses under the GEE. This study signified an improvement on the works of McDaniel et al. [2013] who made a basic assumption of the skewness parameter in the model and the work by Prentice [1976] and Nagler [1994] who propose methods for estimating the value.

By combining assumptions in both works, the study was able to show a flexible way of linking both for flexible parameter estimations under the GEE framework. The said model together with an application to show the superiority of the Skewed Logit (SL) over the conventional Logit was presented.

The other consideration accounted for, was data that exhibited discrete mass at zero and is right skewed. The study proposed an algorithm for modeling clustering while assuming an independent and exchangeable correlation structure. The study extends the work of Swan [2006] who considered the AR (1) correlation structure on modeling rainfall responses. It was shown that the model is insufficient for clustered data and consider a different correlation structure.

The greatest motivation of this work was the modeling flexibility provided by the GEEs.

1.2 Statement of the Problem

It is evident that non-normal responses are common in health research and are gaining favor among researchers. It is also clear that to get more insight on the data, then we have to employ advance but flexible statistical methods.

Non-normality in data can be determined in several ways for a continuous data. For example, a response variable is said to be non-normal when it violates the common normality test such as Andersen Darling test, when it has excess zeroes and when extreme values are reported.

For a binary data, non-normality is present in the response if it is not possible to put a threshold that will result in the correct estimation in the model. For example, when converting a continuous variable into a binary for modelling, it may be difficult to say where the cut off point is. Assuming you are modelling hospital visits, and you have patients having upto 30 visits. In such a case then, if you decide the threshold for moderate visit is 15, then this means that both a patient who visited the facility once, and those who did 15 times have the same 'weight'.

Conventional approaches to modeling some problems have been found insufficient and result in biased parameter estimates. It is therefore important to explore and develop models that will try improve model fit and minimize estimation bias.

Understanding effects of diseases in children is an important aspect for the country for better policy formulation and maintaining a healthy population. It is also the right direction in measuring SDG goal 3 on people wellness. Evaluating diseases in a population is fundamental, in devising ways of addressing them.

However, some diseases such as bacterial vaginosis have been neglected and considered minor, yet they contribute to morbidities and at worse mortalities in the population.

They have been shown to cause morbidities in the early life of an infant. Our aim is to analyze the effect of the same on infants under a HIV setting over time.

It is key to understand hospital access to a population in order to make meaningful policies. The population in Kenya is growing exponentially, yet the number of public health facilities may not be increasing at a rate good enough to accommodate such.

Some of the facilities remain out of reach, while others could be within reach but then don't serve the full purpose because of lack of required physical and personnel resources. What this means is that patients will still be forced to cover longer distances to access care, especially those with complicated cases that need regular check ups. Understanding the predictors for distance to access such facilities will definitely be helpful in formulating alternative policies that can increase access for such care.

Finally, cost spend at the facilities during visit was key to be investigated. Majority of people would not go to hospital for minor illnesses, due to the cost implication. It's clear that most Kenyans are not in the insurance schemes, and in fact only 18% of Kenyans are [GOK, 2014]. What this means is that, most of the households will be required to pay cash at the facility. However, this could discourage access especially on the poor who have very little resources and many needs.

1.3 Justification

From the problems statement, it is clear that addressing the arising issues is critical. Investigating the association of bacterial vaginosis on infant morbidity and maternal complications during birth, is crucial because infant and maternal morbidities plus mortalities in Kenya are among the highest in the world.

There is also need to address access to health facilities and probably find out the major hindrances to the same. This could possibly assist in creating policies that are geared improve to access and encourage the population to seek care at the facilities.

A deeper analysis regarding the cost and the distance will definitely unearth deeper understanding on why Kenyans will chose to attend or not to attend a hospital when they are sick, and further understand the common covariates that can determine access due to cost and also distance.

Although cost and distance have been discussed in the literature as the main determinants of access, it was important to investigate on a larger scale their predictors in order to have a better understanding regarding them.

Such information does not only assist in policy formulation, but also contribute in adding literature on the subject. This will have a positive effect on a further reduction on dependency in the population, and an increase in capital productivity amongst the population within the country.

1.4 Objectives

The general objective was to develop flexible models for analysing Non-normal data in health research.

Specific Objectives

1. To investigate the skewness property of binomial longitudinal data and its application to model infant morbidity under HIV setting
2. To evaluate and develop flexible models with discrete mass at Zero while assuming independence structure under GEE using the Tweedie distribution
3. To develop models for alternative estimation of the scale parameter under the GEE framework
4. To design methods of analyzing clustered data that relaxes the non-normality and correlation assumptions in the response using the Tweedie distributions

1.5 Organization of the Thesis

Chapter 2 is a review of the literature associated with this work, chapter 3 outlines the methodology, chapter 4 are the results and discussions and finally chapter 5 presents conclusion, recommendations and further research.

The readers are advised to follow this sequence 2.2, 3.2 and 4.2; 2.3, 3.3 and 4.3; 2.4,3.4 and 4.4; 2.5, 3.5 and 4.5

Chapter 2

Literature Review

2.1 Introduction to Literature Review

This chapter is about the literature that has been considered in this thesis. We have reviewed literature for each specific objective and the data considered during application.

2.2 Bacterial Vaginosis (BV)-HIV-1 Co-Existence on Maternal and Infant health

Bacterial vaginosis (BV), also referred to as vaginal dysbiosis, a state characterized by altered vaginal biotas has been shown to be a risk factor for birth complications, including low birth weight and preterm births. Current birthing practices are designed to optimize newborn exposure to maternal biota for which she already provides immunologic protection through transplacental immunoglobulin transfer and breastfeeding and thereby

reducing the risk of infection [Fouda et al., 2018], one of the leading causes of deaths among newborns.

A very high prevalence of BV has been described among human immunodeficiency virus (HIV)-infected women. However, data on the effects on child morbidity and mortality remain scarce in Kenya.

Among immuno-suppressed women who are HIV-infected, exposure to BV seems to be more common [Jamieson et al., 2001]. BV is characterized by a lack of *Lactobacillus bifidus* and predominance of anaerobic polybacteria [Lepargneur and Rousseau, 2002, Priestley et al., 1997] such as *Streptococcus*, *Staphylococcus*, *Enterobacteriaceae*; *Candida albicans*; and *Trichomonas*,.

BV is associated with spontaneous abortions and second-trimester miscarriages [Isik et al., 2016], fetal malpresentation, preterm birth [Guaschino et al., 2006, McGregor and French, 2000], postpartum infections, and rupture of membranes [Haggerty et al., 2004, van der Heyden et al., 2013]. Increasing evidence suggests that low birth weight (LBW) and very low birth weight among infants were associated with BV [Thorsen et al., 2006], early neonatal deaths [Ravikumara and Bhat, 1996], as well as compromised immunity.

The prevalence of BV varies from one country to the other and among different races, however, it is more frequent among women in sub-Saharan Africa and women of African ancestry in different parts of the world [Guaschino et al., 2006]. For example, [Alcendor, 2016] investigated health disparities in BV and its implications for HIV-1 acquisition in African-American women and reported prevalence rates of 52% and 32% among Black

and Mexican American women respectively. [Kamga et al., 2019] reported a prevalence rate of 26% among pregnant women in Cameroon.

In comparison, [Nduati et al., 2000] reported a prevalence rate of 47% in this group of Kenyan women who are now the subject of this secondary data analysis Other studies have reported a prevalence rate of 50% for BV among HIV-positive women [Alcaide et al., 2015]. The actual mechanisms underlying BV and the associated risk factors are still poorly understood [Freitas et al., 2017]. It is crucial to note that BV could afflict non-pregnant and pregnant women [Freitas et al., 2017, Kamga et al., 2019] and can also occur in both sexually-active and -inactive young women [Bump and Buesching, 1988], and in young and older women. Still, it is more pronounced among younger women [Dingens et al., 2016].

Current birthing practices are designed to optimize newborn exposure to maternal biota for which she already provides immunologic protection through trans placental immunoglobulin transfer and through breastfeeding and therefore reducing risk of infection, one of the leading causes of newborn deaths. Preliminary data comparing HIV-uninfected and infected women showed very little differences in the vaginal biome of women. However, HIV-exposed sterile babies had different biomes from those of HIV-unexposed babies [Chehoud et al., 2017].

There has also been established a link between HIV transmission and BV [Farquhar et al., 2010, Jamieson et al., 2001] making this a public health concern . Previous research showed that BV modifies the vaginal microbiome, and enhances the transmission of sexually transmitted diseases (STDs. Some studies among HIV-infected women have

revealed that BV is related to an increase in genital shedding of HIV RNA [Sha et al., 2005].

The study hypothesized that BV is correlated with an increase in infections among infants and especially in the context of HIV. This study assessed several determinants of infant morbidity. It was found that the increase in morbidity is related to the development of abnormal microbiota among infants that exposes them to ailments. Simultaneously, maternal HIV infection has been linked to the negligible provision of resistant immunity against ordinary germs among infants [Jallow et al., 2017]. Since any strategy aimed at tackling diseases focuses more on risk factors, it is essential to understand any risk factors associated with BV. Interventions targeting these novel risk factors associated with BV could lead to more effective prevention of morbidities and mortalities affecting mothers and infants.

The risk factors associated with BV in the context of HIV are poorly understood as there have been no reports in these regards. Studies on BV in the context of HIV are few [Atashili et al., 2008, French et al., 2011, Schmid et al., 2000, Spear et al., 2007]. Still, limited results and data have emanated from Kenya regarding the prevalence and associated risk factors among HIV-exposed women. The few studies conducted were cross-sectional surveys and only reported an increased risk of STDs among women exposed to BV and not necessarily the risk factors associated with this combination. Most available literature has focused on the health of infants during and after birth, ignoring that of the mother.

The present study aimed to extend and investigate whether there were any differences in the health of mothers exposed to BV after birth. It has been argued that a healthy

mother could positively impact the proper growth of her child. Several authors have cited the importance of a mother's care and [Nduati et al., 2000] reported higher mortality and morbidity rates among children whose mothers had died. Therefore, maternal health is an important determinant of infant health [Freitas et al., 2017, Keats et al., 2019]. Since this study focuses on infant survival and wellbeing, it is inevitable to consider the aspect of maternal health as the two are interlinked.

Although a few studies have assessed the effect of birth-related complications on the maternal health status after birth, much attention has been accorded to these effects in first world countries. This work focused on a resource-limited country (Kenya) whose HIV prevalence is still high. Understanding the differences in morbidity evolution at different times in the growth of infants, and birth-related complications in a country with limited resources would provide an excellent platform for better policy formulation, planning, and execution.

Data from a longitudinal study to determine risk factors for mother-to-child transmission of HIV among antiretroviral drug naïve women provides an opportunity to determine whether BV increases the risk of early infant mortality and morbidity in this group of HIV-exposed infants. ARV's for prevention of other to child transmission of HIV were introduced in year 2000 in Kenya, well after completion of the data collection phase of this study and therefore all the study participants were ARV naive.

2.3 Non-Normality and Correlation in an Infant Morbidity Longitudinal study

Skewed and non-normal data are commonly observed in health research. Usually, the dataset is transformed, censored, or truncated to impose normality, rather than modeling the data in its natural state [[Manandhar and Nandram, 2019](#)]. Many conventional approaches to modeling lead to incorrect estimates of parameters and standard errors due to the assumptions imposed.

For example, imbalances can occur in binary response data, when symmetry is violated. There are two types of models which are typically employed to analyze data in these scenarios - 1) logit and 2) probit models. Logit models have error variables that follow a logistic distribution and this type of model is considered to be characteristic of discrete choice models. The probit model uses the cumulative standard normal distribution function and assumes the error term is normally distributed. Although this assumption is viewed as a reasonable compromise to achieve mathematical simplicity and parsimonious results, its suitability has been doubted of late.

Several recent studies have investigated various ways of handling non-normal data. However, few have focused on the methodology. For example, a paper published by [[Bono et al., 2017](#)] details several non-normal distributions typical in health, education, and social science, but their substantiation in the literature remains scarce. Further, several other distributions are not considered in this paper, suggesting that they were not common in the study's period of reference. However, these distributions could be vital in

answering some important scientific questions on binary responses that suffer from substantial departure from the commonly assumed symmetric logistic distribution [Nagler, 1994, Prentice, 1976, Tay, 2016]. .

For example, for the Bernoulli distribution, binary asymmetry is defined as the sensitivity to changes in the independent variable that is not maximized at 0.5. This means that a stimulus in any of the independent variables for any individual with probability $P = 0.5$ is not exaggerated. Assuming symmetry in some settings could be inefficient and can lead to biased estimators [Nagler, 1994].

The importance of normality and symmetry in traditional methods of data analysis cannot be under-estimated. There is a need for compromise between statistical simplicity and plausible estimates of parameters when these assumptions do not hold. Questions regarding the suitability of the assumption-based methods have been raised in the literature [Nagler, 1994]. Put differently, although the numerous probability distribution function options can fit the data quite well, the data need to speak for themselves, rather than being forced into a model with assumptions [Manandhar and Nandram, 2019].

There is mounting scientific evidence regarding the inconsistency and weakness of the logit and probit models for skewed binary response data. Recent studies have proposed alternative methods for handling binomial responses, such as: a gamma generated logistic distribution [Castellares et al., 2015], gamma and log-normal distributions [Faddy et al., 2009], improved analysis for skewed continuous responses [Afifi et al., 2007], a skewed Weibull regression model [Caron et al., 2018], a generalized logistic distribution [Rathie et al., 2016], and a skewed logit model [Nagler, 1994]. This shows that modeling non-normality continues to be a topic of importance in recent general research. However,

few methods have been considered and applied in health research. Most of the literature and applications have focused on cross-sectional data in social, political, and economics research [Coelho et al., 2013, Hay et al., 2019, Tay, 2016, Wright et al., 2013, Zhang and Timmermans, 2019].

This study is focused on (1) proposing an improved algorithm for handling asymmetry in binary responses and (2) applying the algorithm in a longitudinal study on infant morbidities.

Morbidity is the state of being symptomatic or unhealthy due to a disease or condition [Hernandez and Kim, 2020] and this can be experienced at any stage in life. This study is focused on BV related morbidities, since this remains a major point of concern globally and particularly in Africa, where the majority of BV cases are recorded [García-Basteiro et al., 2017, Kinney et al., 2010, Tlou et al., 2018]. Child morbidity and mortality as a consequence of BV in conjunction with human immunodeficiency virus (HIV) has been a significant hindrance to meeting goal three of the United Nations Sustainable Development Goals (UN-SDGs) on Good Health and Well-being [Kinshella et al., 2020], which aims to end preventable deaths of newborns and children under 5 years of age.

The scientific literature has established a link between BV and adverse outcomes in mothers and their children [Mwenda et al., 2021b]. Past studies have investigated the occurrence of health deficiencies [Alcaide et al., 2015, Alcendor, 2016], pregnancy loss, labor complications and preterm delivery [Brocklehurst et al., 2013, Carey et al., 2000, Guaschino et al., 2006], as well as spontaneous and recurrent abortions [Isik et al., 2016] among mothers, while others have reported adverse outcomes such as neonatal malformations [Dingens et al., 2016] and low birth weight [Hillier et al., 2018] among

the babies. While some studies have tried to investigate the effects of BV in the context of HIV infection [Alcendor, 2016, Burns et al., 1997, Jamieson et al., 2001], there is still a lack of knowledge regarding the long-term effects in these cases.

To shed light on this topic, this thesis applied the skewed logit model using Generalized Estimating Equations (GEE) to evaluate the variations in the data across time in months and thereby, better understand the infant morbidities. This approach relaxes the strong conditional probability on a binary response, thereby accommodating for the heterogeneity of repeated measures on the same subjects, and accounting for interaction effects in the selected covariates across time.

2.4 Outpatient Care Cost in Kenya

Kenya is classified among the Low middle income countries (LMIC's) and among the fastest growing in Sub-Saharan Africa [Bank, 2015]. To enhance a steady economic growth and proper social development, there is an emerging need to stabilize the National Health systems [Kukla et al., 2017].

The country has continued to strive to reform its healthcare system, but faces challenges such as financial constraints, high debt, weak institutions capacity and large unemployment rates which in turn raises the rate of dependency ratio [Pezzulo et al., 2017], thus huge obstacle to achieve any meaningful change. With constrained budget, the monies allocated for healthcare remains low [Kimathi, 2017], and to achieve any substantial gain, comprehensive improvements or a complete overhaul of the health sector in Kenya needs to be done [Kukla et al., 2017].

Due to the limited resources in LMICs, sound and accurate evidence is needed to make working health policies, which have been found to be more influenced by the current countries economy [Rabarison et al., 2015]. This means that enough data is required to inform strategies by the health professionals in Governments, which are scarce in this context. This would mean that tough choices in developing countries regarding resource allocation and spending, with a view of maximizing the outputs have to be made [Robertson et al., 2019].

In contrast, the developed countries, have continuously benefited from medical security policies, enhanced proper medical care of their citizens, alleviated economic burden of diseases by reducing the catastrophic spending on health and provided financial support to ease the health burden all this made possible, by availability of current data [Jing et al., 2020, Kato and Okada, 2019, Lee and Shaw, 2014, Li et al., 2019a,b, Liu and Dai, 2020].

In the next decade, literature asserts that the demand for in and outpatient healthcare is likely to increase [Group, 2020] and this exerts pressure on governments of developing countries with limited resources on how to handle the expected increase/expansion surge in demand. Research has established that more than 11 million Africans, within which 0.45 million Kenyans are pushed into extreme poverty each year because of out-of-pocket health expenses for both in and outpatient.

To caution its citizens against health care spending strain, there have been consistent efforts by the Kenyan government to have most of its population insured through the National Health Insurance fund (NHIF), however, 83% of the population remains uninsured [Barasa et al., 2017]. Efforts to reform the fund [Mbau et al., 2020] which could

be a gateway of achieving the universal health care(UHC) for all in Kenya were established [Barasa et al., 2018b].

This was achieved by selection of a few counties (Nyeri, Kisumu, Machakos and Isiolo) that were to act as pilot in which the state was to meet all the medical costs [Obare et al., 2014, Okech and Lelegwe, 2015] and enhance achievement of the SDG goal 3 [Barasa et al., 2018a]. The outcome of the pilot was to inform a possibility and sustainability of rolling out the whole program-me to the whole country.

This is inline with the constant global push towards UHC in LMIC's, and this has necessitated reforms on health sectors to try and achieve this. The main objective of the UHC was to caution citizens against the catastrophic and impoverishing effects of out-of-pocket healthcare payments in Kenya [Chuma and Maina, 2012, Salari et al., 2019] that has led to poverty in households [Kimani, 2014], Socio-economic inequality and inequity in use of health care services [Ilinca et al., 2019], and lots of time wasted and longer distance travelled to access healthcare services [Kukla et al., 2017].

Unfortunately, measured on global level Kenya's strides remain very inadequate [Obare et al., 2014] and feedback from the pilot project was that the UHC programme could not be fully supported by the Exchequer.

Household spending on outpatient care is a very important characteristic of measuring people's health in terms of finances to sustain a good health. However, it has not received attention on recent literature due to the un-seriousness of health conditions that pose no danger. However, it is important to note that some health conditions which appear

insignificant, can easily deteriorate with time if not given proper medical attention, thus it is very key to arrest the situation at the outpatient level.

The choice of seeking outpatient care when sick or injured, could be influenced by (1) the seriousness of the health conditions of the said individual and (2) the financial ability to pay for the required service [Awiti, 2014]. In this case, household characteristic of the member of the households or the care provider from which the member of the said household needs to seek help is very key [Ensor and Cooper, 2004, Mwabu, 1989, Umar et al., 2012].

Most of the decisions are determined by the head of the household, who mostly acts as the bread winner make vital decisions of any given household mostly on how resources are spent [Posel, 2001].

Therefore, this motivated this study to investigate the outpatient care predictors in reference to several characteristics of the head of household.

2.5 Distance Traveled for Inpatient Care in Kenya

Inpatient care is defined as a case in which an individual is hospitalized for more than 24 hours and reflects a more serious health complaint. An estimated 1.2 million Kenyans required these services in 2013, and the number is predicted to increase exponentially in the upcoming decade [GOK, 2014]. Among those seeking care, various factors are key in predicting distance traveled.

For example, differences in wealth would determine the distance traveled. Additionally, those in the higher wealth quintiles can afford to pay fees in any facility, pay insurance premiums that can guarantee admission at any facility, and have the financial strength to pay cash at the given facility of choice.

In contrast, those in the lower-income quintiles have fewer choices of the type of facility for care, as they are limited by finances. Also, although government healthcare facilities are much more affordable and are the best choice for care, most are miles away and out of reach from places of residence.

To mitigate this, the Kenyan government has increased the establishment of as many inpatient services as possible, which includes the upgrading of healthcare facilities that currently only offer outpatient services by equipping them with machines that are needed to provide inpatient care. However, this effort often requires the provision of essential services such as water and electricity, accessible roads, and housing, which limits some facilities for upgrades.

Some of these facilities are found in rural areas and slums, which serve a large number of people, meaning their upgrade would be significantly beneficial for the residents. However, currently, the poor who live in these areas continue experiencing difficulty when they need inpatient care. Thus, ease-of-access policies for inpatient services should be implemented.

Distance to inpatient services can determine the well-being of a population and is potentially linked to individual survival. For example, there has been a link between long distances and poor health outcomes, including longer lengths of stay in hospitals, nonattendance at follow-ups [[Kelly et al., 2016](#)], and, at the worst, patient fatality [[Karra et al.,](#)

2017].

A study conducted in Zambia found that long distances and the lack of geographic access to much-needed obstetric care for pregnant mothers explain why there are still fatalities due to deliveries without skilled care [Gabrysch et al., 2011], and, in Tanzania, child mortality has increased due to the lack of access to healthcare facilities [Kadobera et al., 2012].

In contrast, a short distance to a facility has been associated with higher utilization of the facility and better health outcomes in sub-Saharan Africa [Schoeps et al., 2011]. In the event of an emergency, distance could be a defining factor for patient survival, with long distances predicting higher mortalities [Kadobera et al., 2012]. Studies across some developing countries, such as Bangladesh [Biswas and Kabir, 2017], Kenya [Escamilla et al., 2018], Nigeria [Awoyemi et al., 2011, Stock, 1983], Afghanistan [Nic Carthaigh et al., 2014], and Burkina Faso [Schoeps et al., 2011], point out the importance of distance in predicting health outcomes.

However, although a correlation between distance and decay exists, with those further away associated with underuse and those closer associated with appropriate use, there is little evidence to show how this translates to health outcomes. Therefore, distance traveled to acquire the required inpatient services requires further investigation and thus forms the basis of this paper.

This study focused on the secondary analysis of the Kenya Household Health Expenditure and Utilization Survey (KHHEUS), collected in 2018, and centers on question Q68: *What distance did < name > cover in kms to get to the inpatient facility?*

This study builds on non-normal response analyses under generalized estimating equations (GEE) by [Mwenda et al., 2021a] and adopts the approach of [Swan, 2006] and [Kurz, 2017]. [Kurz, 2017] analyzed healthcare utilization cost data using a Tweedie distribution, but his works were based on a generalized linear model, meaning correlation was not considered. In contrast, Taryn and Mwenda et al. [12] considered a decaying correlation with time but with applications to rainfall and health data, respectively.

Data used in this study was based on clusters that exhibit patient-to-patient correlation characteristics, meaning if this work used the previous methods (e.g., correlation decay), then will have incorrect results. Therefore, due to the clustered correlation nature of data used within counties, established a new approach to model the response using the Tweedie distribution by considering what this work refer to as decay distance with constant correlation and uses the exchangeable correlation structure under a GEE framework.

The main goal of the study was to identify which covariates were best associated with distance traveled for inpatient care in Kenya, which were obtained in this study. To carry out this kind of analysis, this study created an R function with a Tweedie distribution and exchangeable correlation structure under the GEE framework. Due to difficulty in linking inpatient admissions and accessibility, this work relies on self-reports from respondents on the distance they covered to access the healthcare facility.

Statistical literature review on the Tweedie

Tweedie distributions have been widely applied in modeling non-normal response data with a discrete mass at zero because they can incorporate skewness without any data transformation. Most of the methods suggested in the literature for the analysis of such data mainly consider data transformation [Manikandan, 2010], two-way analysis [Su et al., 2018], and Bayesian methods [Swallow et al., 2016].

However, these methods are not efficient for our approach because of the correlation nature of our data within clusters. Approaches proposed for analyzing non-normal data in the generalized linear model (GLM) framework have the limitation of ignoring the correlation, which may exist among subjects who belong to the same cluster. Moreover, the methods require the specification of a full likelihood. This means that if the likelihood is misspecified, the results will be incorrect.

The method used in this study, which used quasi-likelihood methods, only requires specifying how the mean relates to the covariates. It is also very flexible in that, in the event of the misspecification of the correlation structure, the estimates are still plausible. Moreover, the methods developed by this study are easy to modify and adapt.

Previous evidence suggests the influence of covariates on distance to healthcare facilities; however, a research gap on the selection of the best-fitting covariates remains. This study aims to determine the combination of covariates that influence the distance a Kenyan citizen will travel to seek inpatient care. This new work enhances the application of the Tweedie distribution to understand the influence of a given set of covariates on distance.

The Kenya House Hold Expenditure and Utilization Survey (KHHEUS) inpatient data were obtained, cleaned, and coded. Thirteen (13) covariates for the dependent variable, distance were investigated. Residence type was categorized as rural or urban. Five wealth index quintiles ranging from richest to poorest were constructed from the ownership of different household assets using the principal component analysis as described by [Filmer and Pritchett, 2001].

The education categorization followed the justification and methods provided by [Rippin et al., 2020]. This work considered four categories: those who never went to school (those under 3 years of age and those who responded, "Never went to school"), lower education (pre-primary, primary, and informal [madrassa]), intermediate education (secondary, vocational, and college), and higher education (university degree or higher).

Age groupings for employment followed those defined by the Organisation for Economic Co-operation and Development (OECD). This work divided the patients into four age groups: those aged below 15 and above 65, who are considered unable to work; age group 15–24, who are entering the labor market following an education; age group 25–54, who are those in their prime working lives, and age group 55–64, who are passing the peak of their career and approaching retirement [OECD, 2020].

This thesis divided household size into three categories: small (1–3 members), medium (4–6 members), and large (7+ members). It also categorized the healthcare admission duration into 1–5 days, 6–20 days, and 21 or more days. Access to insurance was classified as yes or no.

The head of a household was classified as male or female. Amounts paid for healthcare

were categorized into three groups: low (1–3,000 KES), medium (3,001–10,000 KES), and high (10,000+ KES). The dependent variable, distance, was assumed to be 0 km for any value captured and less than 2 km, following other studies in Kenya [[Mwaliko et al., 2014](#)].

Chapter 3

Methodology

3.1 Introduction

This chapter considers the methodologies adapted for this work

3.2 Suppressor Effect Application for Modelling Correlation under Independence Assumption

This study is based on a 25 year old data set of the randomized trial of breastfeeding and formula among HIV infected women. [Nduati et al., 2000] embarked on a randomized clinical trial of breastfeeding versus formula among HIV infected women to enable determination of the risk of breastmilk transmission, correlates of transmission and maternal outcomes.

Women were enrolled during pregnancy and a careful assessment was made among others the presence of sexually transmitted infections including bacterial vaginosis to facilitate determination of factors associated with mother-to-child transmission of HIV. Women were then followed through delivery and there after mother-baby pairs were followed up into end of the study at 24 months post-delivery or the death of an infant.

3.2.1 Study population, enrollment, delivery, and follow up

This study used data on a study that was conducted in Nairobi with active enrolment from 6th November 1992 to 7th October 1997. Sixteen thousand five hundred and twenty-nine (16,529) women attending 4 antenatal clinics were screened for HIV, 2315 were found to be HIV positive while 14,214 wer HIV negative. Of the HIV positive, 1708 returned for results while 607 did not. Of the women who returned for the results, 425 met study criteria for enrolment.

Of the 425 eligible and screened for BV, our interest lied on the 401 who had life singleton births after excluding still births, miscarriages maternal deaths and those lost to follow up. Enrolled women were subjected to a standard interview and physical examination at each prenatal visit, delivery and postnatal visits at 6, 10, 14 weeks and then monthly until the child was one year and thereafter every 3 months until 24 months or death of an infant.

Maternal enrollments and follow-up: During enrollment a physical examination including a pelvic examination were performed. During the speculum examination vaginal and cervical secretions and samples were collected for microscopy and gram staining for bacterial

vaginosis.

At delivery a standard form was used to collect delivery data. Women were encouraged to deliver at Kenyatta National Hospital and study team nurse midwives provided 24 hour cover to facilitate this process. To determine the viral load and CD4–8 cell counts, 15 ml of blood was collected in purple-top vacutainers. After delivery, blood was drawn from each infant for testing, and the mother and infant pair was followed up every month in the first year. At every visit, a history was obtained from the mother and the pair underwent physical examination using standard tools. At the time of delivery, after excluding stillbirths and second-born twins, 401 dyad pairs remained. During birth, 348 pairs of mothers and infants were available for analysis, 169 mothers were exposed to BV while 179 were unexposed. This is after excluding Fifty-three incomplete cases. At the end of year 1, only data regarding 328 pairs were available for analysis. This is after 20 pairs were excluded from further analysis for the following reasons: 14 babies died and their morbidity measures were not assessed thereafter and 6 mothers were lost to follow up. These were finally included in the multiple logistic regression analysis. Among them, 157 mothers tested positive for BV, while the remaining 171 were negative. Flow chart shown in (Figure 3.1).

Subsequently, on the final pair included, the prevalence of BV was calculated as $(159/328)=48\%$, which is similar to what [Nduati et al., 2000] reported, and consistent with other studies that put the generality in Kenya to be between 30-50% [Cohen et al., 2012].

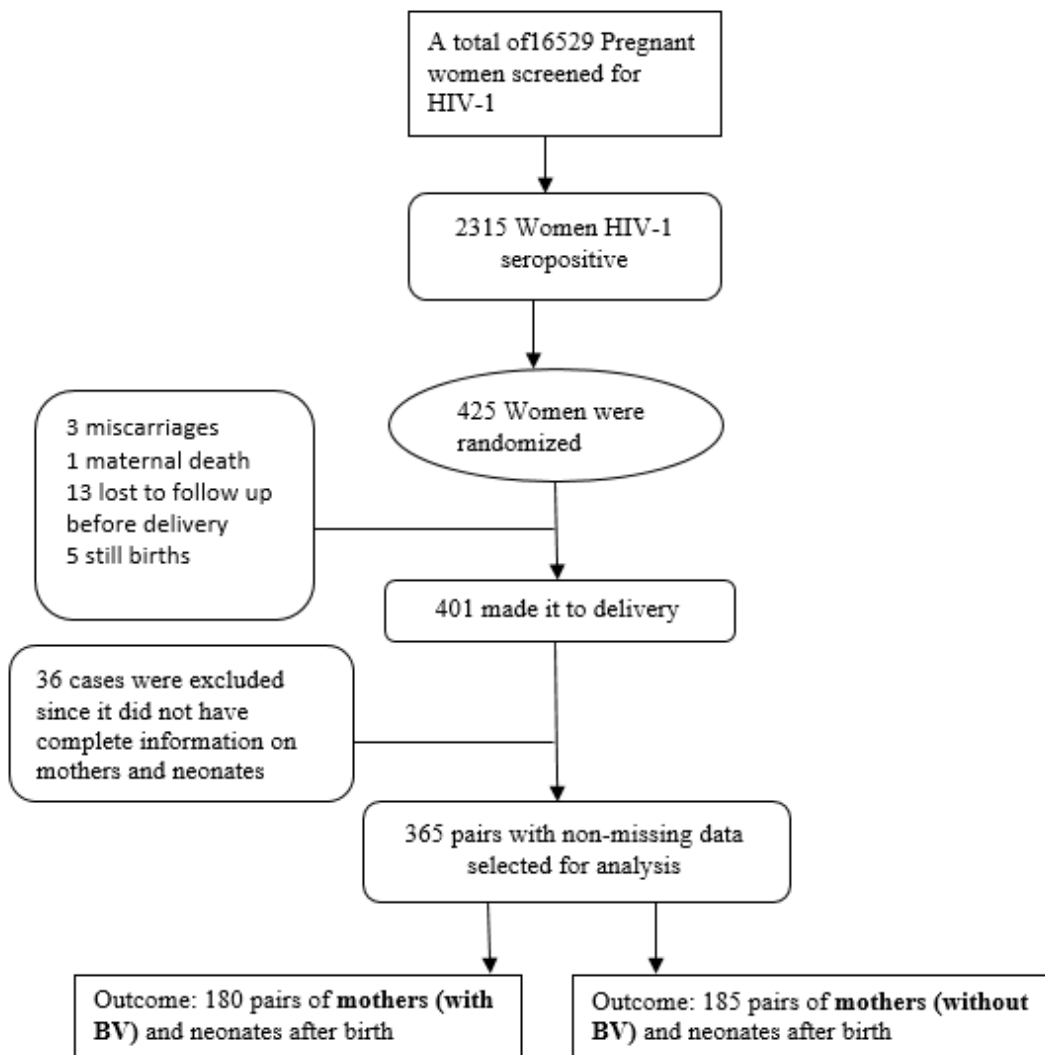


FIGURE 3.1: Trial flow diagram describing the recruitment of cases and exclusions

3.2.2 Clinical characteristics

The incidences of morbidities were both self-reported and recorded from hospital visits. A clinical assessment of clinical characteristics were carried out using a standard tool at scheduled study visits and non-scheduled ones due to illness.. The occurrence of each illness was recorded for every infant at each visit.

The process included documentation of any illness or hospital visits/admissions in the period since last clinic visit. At the visit there was interview using a standard tool for

any symptoms of illness, a physical examination and collection of relevant study samples by the team of study doctors, an obstetrics registrar and three consultant of whom two were pediatrician, and one medicine-pediatrics specialist.

During birth and immediately after birth, several clinical characteristics were assessed for both neonates and mothers. Among mothers, we considered signs that were indicative of complications, including excessive bleeding, urinary tract infection, and hypertension.

Infant assessment at birth: Among neonates, in-hospital characteristics were classified as admissions lasting longer than 24 hours (classified as yes or no) and the number of days spent at the hospital. Other characteristics of neonates were assessed by measuring the length (cm) after birth, head circumference (cm), and weight (g). Apgar score at 1 and 5 was recorded at delivery while infant gestational age was carried out using the Dubowitz scores within 7 days of delivery and those scored below 37 weeks of gestation were classified as premature. Any issues of breathing were classified as respiratory distress. We also assessed the neonates for jaundice, conjunctivitis, lymphadenopathy, skin rash, Asphyxia, Pneumonia, Sepsis/meningitis and other abnormality.

Infant assessment during follow-up: At every scheduled and unscheduled visit the babies underwent anthropometric measurements (weight, height, and head circumference) and a clinical assessment for the presence of common childhood morbidities, including but not limited to pneumonia, ear infection, Blood in stool, lymphadenopathy, encephalopathy, sepsis, conjunctivitis, dehydration, wheezing, hematologic conditions, cold, otitis, fever, cough, diarrhea, thrush, vomiting, difficulty in feeding, heat rash, fungal rash, Eczema/-dermatitis, Scabies and mouth ulcer. In order to capture all morbidity events, the participants had unrestricted access to care in the study clinic for that run 4 days a week and

had careful instructions of how to navigate the hospital for other services outside clinic hours.

3.2.3 Laboratory methods for detection of Bacterial vaginosis

Details on how other specimens were obtained have been described elsewhere [Nduati et al., 2000]. Test specimens for BV and HIV-1 were collected via pelvic speculum examination. Vaginal and cervical specimens were collected separately using sterile Dacron swab. The genital infections, including BV, were diagnosed and treated before delivery. Women were categorized as having BV using the Nugent criterion (a pH of ≥ 7 in the specimen was considered significant).

3.2.4 Ethical Approval

The study protocol was approved by the ethics review boards of the University of Washington and the University of Nairobi.

3.2.5 Statistical analysis

The effects of BV on neonates and mothers in the Nairobi study were described using ORs. Continuous variables were reported as means and standard deviations while categorical variables were reported as frequencies and proportions.

We subsequently analyzed the morbidity incidences ever reported by the infants between the two groups using Pearson's chi-squared test and computed the p-values using the

fishers exact text. However, only 2 morbidities seemed to show any association with the BV. We then employed the multiple logistic regression modeling approach under the Generalized Estimating Equations, using the independent correlation structure.

The inclusion of all the morbidities in the model was supported by the fact that some morbidities which were not significant or associated with BV were significant in the multiple logistic regression. Additionnaly, we adapted the suppressor effect concept proposed by [Sun et al., 1996] and argue that the methodological approach could bring out some patterns of the disease which the conventional approaches could fail to reveal.

Finally, to assess the effects of BV on survival, we estimated the cumulative hazard using the Kaplan–Meier method, to identify any differences between the two groups. The method is attractive due to its ease in interpretation of the data such that a researcher can tell the probability of an infant dying on condition that their mother had BV during pregnancy and birth.

All statistical analyses were performed using R version 3.6.3 (R Development Core Team, Vienna, Austria) [R Core Team, 2017]. Analysis items with $P < 0.05$ were considered statistically significant.

3.3 Skewed Logit model for correlated binomial longitudinal data and application to modelling infant morbidity under HIV setting

3.3.1 Materials and methods

Data

The Nairobi Infant Morbidity Study (NIMS) was a randomized clinical trial carried out by scholars in the International AIDS Research and Training Program supported by grant NICHD-23412 from the National Institutes of Health. The objective was to collect high quality longitudinal data on morbidity and mortality of babies from HIV-positive pregnant women in a random sample considering mothers who either breastfed or gave their baby formula. The description, analysis and findings of the original study can be found elsewhere [[Nduati et al., 2000](#)].

The study participants were drawn from a population of 16529 pregnant mothers attending four antenatal clinics in Nairobi, Kenya. After screening for HIV, 2315 were found to be positive. Of these women, 425 were selected and verbally agreed to be enrolled in the study. At each prenatal visit, each woman was subjected to a standard physical and clinical examination, and an interview.

Before birth, at 32 weeks of pregnancy, pelvic examination, including analysis of vaginal and cervical secretions were conducted for each woman to determine their BV status. This was done using sterile Dacron swabs by a trained clinical officer and the Nugent

criteria was used to qualify a woman for a BV diagnosis. A pH value from the swab, of ≥ 7 was considered a case, indicating alkalinity of the vaginal fluids and inhibition of bad bacteria such as as *Trichomonas*, *Candida albicans*, *Enterobacteriaceae*, *Staphylococcus* and *Streptococcus*

Immediately after birth, infants were assessed for HIV using enzyme-linked immunosorbent assay (ELISA). Those who tested positive were subjected to a more accurate Polymerase chain reaction (PCR) test. Infants who had three consecutive negative tests were deemed negative. The pairs of infants who survived were regularly re-examined over the next two years and their history of ailments were documented at every visit.

The study data was collected in two ways, scheduled and unscheduled visits. Scheduled visits meant that the dyad pairs were supposed to come to the clinic for examination at a specific time, while unscheduled visits meant they could pop in any time in case of an illness. Other physical examinations of the baby, including details like sex, weight, and height, were observed and recorded

The planned visits were bi-weekly during the first 3 months and monthly thereafter for up to two years. In all scenarios, data were collected either through parental report or diagnosis at the hospital or clinic. Of the total number of women enrolled, complete records from birth to six months were only available for 401 women. The other 24 women either had miscarriages or still births or did not complete the follow up appointments. Of the 401 women, 74 pairs had missing values, either for the mother's BV measure or for the morbidity incidences of the infants. To address this, we applied a missing completely at random (MCAR) mechanism. There is sufficient evidence that, using the GEE approach,

this approach still enables a consistent estimate of the regression parameters so long as the mean model is correctly specified [Laird, 1988]

A standard questionnaire developed by the principal investigator of the study to identify illnesses was completed for both the mother and the child. This was achieved using a 19-item yes/no morbidity questionnaire which purports to measure health status of an infant. The total score of the questionnaire is computed as the count of all the "yes" responses. There were a total of 1962 observations from 327 pairs of mothers and babies. From the total score, we created a binary response of; (1) those who did not have any illness and (2) those who had either minimal or severe illnesses. Table 4.9 presents an initial exploratory analysis used to identify the asymmetry in the total responses for each month. This evidence of asymmetry justifies the use of the skewed logit model.

Ethical approval

The study protocol was approved by the institutional review boards of the University of Washington and University of Nairobi. Verbal consent was obtained from all mothers prior to their inclusion in the study. The investigators in the study did not require documentation of any consent for the participants because at the time of the study, written consent was not mandated by the ethics bodies involved. Therefore, at that time, no procedures regarding written consent were violated given the research context of doing the study in Kenya.

3.3.2 Statistical Model

Generalized Linear Models (GLMs) and the GEEs were used to model infant morbidity. The first modeling approach to determine the need for the skewed logit model and the value for skewness was carried out with GLMs. The response variable was the health status of the infant within a particular month at the time of the hospital visit or the reported health status about the infant from the mother. For our study, we considered all health events, whereby a health event occurred if an infant was reported to have experienced any illness within the month.

Let Y_{it} be the response for subject i measured at different points in time $t = 1, \dots, n_i$ denote the outcome vector for subjects $i = 1, 2, \dots, N$ and \mathbf{x}_{it} is a $n_i \times q$ matrix of covariate variables for subject i . The expected value is given by $E(Y_{it}) = \pi_{it}$ and the linear predictor that relates the mean to the covariates is given by

$$g(\pi_{it}) = \eta_{it} = \mathbf{x}_{it}^\top \boldsymbol{\beta} \quad (3.1)$$

where \mathbf{x}_{it} is the covariate vector for subject i at time t with length q . This includes the infant weight, mother's BV status, HIV status of the infant, and feeding status of the infant. $g^{-1}(\cdot)$ is a known link function such as the skewed logit model and $\boldsymbol{\beta}$ are regression parameters.

For each infant status of illness at any chosen time point, the response follows a Bernoulli distribution with p_i (probability of being ill= π_{it}) and is specified as:

$$Y_{it} \sim \text{Bern}(\pi_{it}) \quad (3.2)$$

To model the outcome, the logit and probit models are preferred options, but they both have conditional probability distributions, which have a maximum at 0, such that P_i for $i \in (0, 1)$ is 0.5 and thus, they have a fixed symmetry of 0.5. However, this assumption of symmetry may not be realistic to all Bernoulli responses and therefore, not desired [Coelho et al., 2013, Golet, 2014, Nagler, 1994]. For this reason, the skewed logit approach is employed here, taking advantage of the fact that the logit model is nested within the skewed logit model as shown in Fig 3.2 . There are reported similarities in terms of model specification, estimation, and iterations. Using the skewed logit model made it possible to see if the data were skewed and therefore, to estimate the skewness value.

The probability of a child experiencing illness is given by

$$\Pr(\text{illness} = 1) = g^{-1}(\mathbf{x}_{it}^T \boldsymbol{\beta}) \quad (3.3)$$

In this work, we aim to consider a response that violates the symmetry assumption, using the framework described above. Following [Burr, 1942], $k^{-1}(\cdot)$ accommodates asymmetry through;

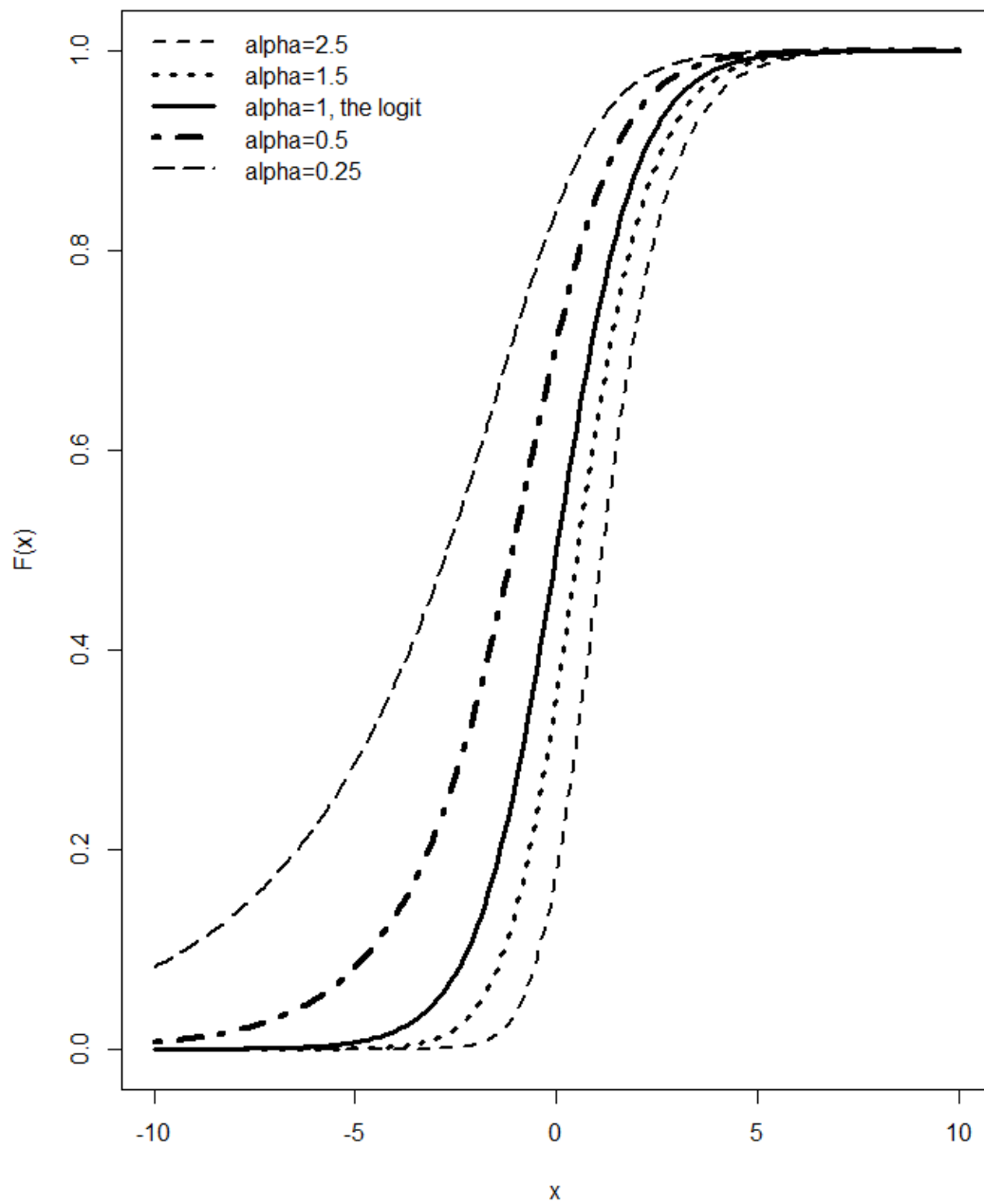


FIGURE 3.2: Cumulative density function of the skewed logit model with different values of skewness. The bold continuous line represents the logit model which assumes symmetry.

$$g^{-1}(\cdot) = \Pr(y_{it} = 1 \mid \mathbf{x}_{it}) = 1 - \frac{1}{(1 + \exp(\mathbf{x}_{it}\boldsymbol{\beta}))^\alpha} \quad (3.4)$$

for $\alpha > 0$ and this is the skew value to be estimated.

This variation implies that the maximum is no longer restricted to $P = 0.5$. Since the skew value cannot be observed, a regression model was fitted of all covariates under the skewed logit model using the GLM approach. Further, the α obtained was used as a proxy for the disturbance to be used in the GEE.

Detailed methodology on obtaining α can be found in Appendix 9

To obtain robust standard errors that are meaningful for the parameter estimates, we adopted the Huber sandwich estimator [Freedman, 2006, Huber, 1964], which has the ability to relax the intra-group correlation. To increase the efficiency of model convergence, we specify a tolerance value of 0.0001 and set the maximum number of iterations to 100.

The applicability of the two models using the set of covariates was determined by the likelihood ratio test that compares the logit and the skewed logit model to identify any significant differences [StataCorp, 2015].

3.3.3 Estimation of parameters using the GEE

Developed by [Liang and Zeger, 1986] in their land mark paper, GEE can be used to model correlated data and give a marginal inference interpretation. The strength of

this approach is its straightforward application, since the mean response depends on the covariates and not on any random effects or any previous responses. Thus, only the marginal distribution of the subject dependent vector is specified.

The variance of the response is a function of the mean and is conditional on the vector of covariates represented as

$$\text{Var}(y_{it} \mid \mathbf{x}_{it}) = v(\pi_{it})\phi$$

where v is the variance function depending on y_{it} and ϕ is the dispersion parameter assumed to be 1 for the exponential dispersion model family.

Let D be a diagonal matrix of derivatives $\partial\pi_i/\partial\eta_i$ and $V(\boldsymbol{\pi}_i)$ is a $n_i \times n_i$ diagonal matrix to be decomposed as;

$$V(\boldsymbol{\pi}_i) = \mathbf{D}[V(\pi_{it})]^{\frac{1}{2}} \mathbf{I}_{(n_i \times n_i)} \mathbf{D}[V(\pi_{it})]^{\frac{1}{2}} \quad (3.5)$$

This estimation equation treats each observation within a given time point as independent. This study focused on the marginal distribution of the response for which the mean and the variance are averaged over the six observation time points. However, the variance of correlated data does not have a diagonal form and hence, we replace the identity matrix $\mathbf{I}_{(n_i \times n_i)}$ using methods proposed by [Liang and Zeger, 1986] with another correlation structure $\mathbf{R}_i(\boldsymbol{\rho})$. G_i is the diagonal matrix with j^{th} the diagonal element equal to $v(\pi_{ij})$ such that equation 3.5 corresponds to 3.6 as shown:

$$W_i = \mathbf{G}_i^{\frac{1}{2}} \mathbf{R}_i(\boldsymbol{\rho}) \mathbf{G}_i^{\frac{1}{2}} \quad (3.6)$$

The working correlation structure R_i with dimension $n_i \times n_i$ is assumed to depend on a vector of the association parameter ρ . [Liang and Zeger, 1986] stated that the mis-specification of $\mathbf{R}_i(\boldsymbol{\rho})$ only affects the efficiency of the $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$ is robust against mis-specification. This study considered several correlation structures. These include the unstructured structure, where every measure between two points is assigned its association parameter; the auto-regressive (AR-1) structure with $lag = 1$, in which correlation decreases exponentially with the differences in measurements; the independence structure in which we use the identity matrix as the correlation structure; and the exchangeable structure in which correlation is assumed to be equal across different measurements. Liang and Zeger have provided evidence that mis-specification of the correlation structure only affects β 's efficiency. This is because of the assumption that the estimation equation for the regression coefficients is orthogonal to the estimation equation for the correlation coefficients.

The GEE are as follows;

$$\sum_{i=1}^j \mathbf{D}_i^T \mathbf{W}_i^{-1} (y_i - \pi_i) = \mathbf{0} \quad (3.7)$$

Where $D_i = \mathbf{G}\Lambda_i\mathbf{x}_i$, $W_i = V(\pi_{it})^{\frac{1}{2}}\mathbf{R}_i(\boldsymbol{\rho})V(\pi_{it})^{\frac{1}{2}}$ and Λ_i is a diagonal matrix with j^{th} entry given by $\frac{dk^{-1}(\eta_{ij})}{d\eta_{ij}}$

The most traditional way of solving the estimating equations is to employ the iterative re-weighted least squares algorithm, which is a modification of the Newton–Raphson algorithm. In this approach, the observed Hessian matrix replaces the expected Hessian matrix, using the Fisher scoring algorithm.

However, [McDaniel et al., 2013] proposed an alternative approach to estimate β 's such that instead of the summation in Equation 3.7, they are evaluated using the matrix form as shown;

$$\mathbf{x}^\top \Lambda \mathbf{G} \left(\mathbf{G}^{\frac{1}{2}} \mathbf{R}_i(\boldsymbol{\rho}) \mathbf{G}^{\frac{1}{2}} \right)^{-1} \mathbf{Z} \quad (3.8)$$

Several methods of analyzing skewed binary data have been proposed in the literature [Bazán et al., 2010, Caron et al., 2018, Chen et al., 1999, 2001]. Of particular importance for this current study is the method described by [Prentice, 1976] that allows for the elimination of asymmetry through the modification of the inverse link function of the logit model, given as:

$$\left(\frac{\exp^{\mathbf{x}^\top \boldsymbol{\beta}}}{1 + \exp^{\mathbf{x}^\top \boldsymbol{\beta}}} \right)^\alpha$$

Statistical analysis was then implemented in **R** version 3.6.3 [R Core Team, 2017] (The R Development Core Team, Vienna, Austria). Though most functions are available directly

in the software, we required an extra library including "dplyr" [Wickham et al., 2020] for data manipulation, "glogis" [Zeileis and Windberger, 2018] for the skewed logit Cumulative Density Functions (CDFs) plots with different values of α and the "geeM" [McDaniel et al., 2013] for the skewed logit analysis under the GEE.

The final GEE models were calculated and the probabilities of a child having morbidities were interpreted. These probabilities were calculated using the inverse-logit function and odds ratio as the exponential values of the differences in the logits. The p -values were calculated for each parameter estimate, as were the Z statistic and the model and robust standard errors.

3.4 Tweedie Distribution for a Response Exhibiting is continuous, non-negative and Right Skewed Characteristic Using Independent Structure with application to health cost data

3.4.1 Exploratory Data Analysis

In a secondary analysis, we used data from the Kenya Household Health Utilization Survey 2018 (KHHEUS 2018). This was a survey conducted to establish costs incurred by households and individuals in Kenya on inpatient and outpatient care among other outcomes from a random representative sample representing all the households in Kenya.

Our outcome variable was the total cost incurred for outpatient care. This consisted of registration card, medicine/chemotherapy/vaccine, consultation, diagnosis tests (x-rays, lab etc.), medical checkup and dialysis. Respondents were asked for the breakdown of each as incurred, and those who couldn't recall the breakdown, there was a column for 'total' where the total cost was captured. This therefore means any inconsistency in recall would not affect our modeling.

In order to establish an association of total cost for the outpatient with covariates, we selected covariates which are commonly considered to predict health care cost and utilization. Among our covariates of choice was, age, gender, level of education, employment, marital status, whether in the household there was a smoker, or any member suffering from HIV, asthma, respiratory problems. Employment was used as a proxy to estimate income of the household head.

From the choice of variables, we only focused on the 1st visit, since there were respondents who reported more than 4 visits, and we focused on households headed by 18 years and above. We summed up the total expenditure for health in the households as we are interested in estimating utilization per household and not individual. After summing up the total costs incurred by the households, and establishing households' heads who were 18 years and over, we got a total of (N=11130). We looked at the respondent for relationship, and if the respondent was not the head, then the second person with closest relationship to the head was re-coded to be the head.

Upon exploratory analysis of our data as shown by Table 4.13. There are 37.01% zero observations on the total spent by a household for outpatient care with a mean of 1141.18

and a standard deviation of 3232.73. Positive cost alone has a mean of 1811.63 and a standard deviation of 3921.27. Clearly both data are skewed the right.

We therefore adapted methods by [Hardin, 2013] that will enable us to check model fit using the QICu criteria. Since QICu is quasi likelihood, we therefore adopted the GEE framework for our model while assuming an independence correlation structure.

We compared our models based on the QICu computed, and reported the results as marginal effects, evaluated as the mean of any given covariates. Finally, we compared the logarithmic and default canonical link to show the advantage of using the logarithmic link. We also plotted the means per age group of the household head to establish any relationship.

3.4.2 Tweedie Distribution

Tweedie distribution are members of the Exponential dispersion model (EDM) whose probability density function can be expressed as

$$p(y; \theta, \phi) = a_p(y, \phi) \exp \left\{ \frac{1}{\phi} [y\theta - \kappa(\theta)] \right\} \quad (3.9)$$

Or

$$p(y; \theta, \phi) = b_p(y, \phi) \exp \left\{ \frac{-d(y, \mu)}{2\phi} \right\} \quad (3.10)$$

where $\phi > 1$ is the dispersion parameter, $\mu = \kappa'(\theta)$ is the position parameter and θ is the canonical parameter and y is the variable of interest. The mean of the tweedie is expressed as

$$E[y] = \mu = \kappa(\theta) \quad (3.11)$$

and the variance given by

$$\text{var}[y] = \phi\kappa''(\theta) \quad (3.12)$$

A response that an Exponential Dispersion Models (EDM) with mean μ and dispersion parameter ϕ such that

$$y \sim \text{EDM}(\mu_i, \phi/w_i) \quad (3.13)$$

where w_i are known weights usually assumed to be 1 and

$$g(\mu_i) = \eta = x_i^T \beta \quad (3.14)$$

Distributions of selected members of Exponential families is shown by Table 3.1 while the distribution for the Quasi-likelihood are given in table 3.2

TABLE 3.1: *Distributions of selected members of Exponential families*

Distribution	$\kappa(\theta)$	$\mu = E(Y)$	Variance function
Normal	$\frac{\theta^2}{2}$	θ	1
Poisson	e^θ	e^θ	μ
Binomial	$\ln(1+e^\theta)$	$\frac{e^\theta}{1+e^\theta}$	$\mu(1-\mu)$
Gamma	$-\ln(-\theta)$	$-\frac{1}{\theta}$	μ^2
Inverse Normal	$-(-2\theta)^{1/2}$	-2θ	μ^3
Tweedie	$\frac{\theta(1-p)^{\frac{2-p}{1-p}}}{2-p}$ for $p \neq (1,2)$	$\kappa'(\theta)$	μ^p for $p \neq (0,1)$

3.4.3 Mean Function

Mean is the first moments and is a function of variance

3.4.4 Variance function

The variance function uniquely identifies a distribution within the class of EDMs whereby the variance is related to the mean through $\kappa'(\theta) = \mu$ and $v(\mu) = \kappa''(\theta)$. The variance function describes the mean-variance relationship of a distribution when the dispersion parameter is constant.

Our interest lies on the last distribution in table 3.1 which is a three parameter family of distribution in the μ (mean), $\phi > 0$ (dispersion) and p (index parameter) that determines the shape of the tweedie distribution. Most of the common distribution are within the tweedie, and all that one needs is to specify the index parameter. for example, when the index parameter is 0(Normal), 1(poisson), 2(gamma) and 3(inverse normal).

The variance is given as

$$V(\mu) = \mu^p$$

where $p \in (-\infty, 0] \cup [1, \infty)$ is the index determining the distribution [Cook, 2000].

Tweedie models exist for all values of p outside the interval (0,1) Apart from the 4 distributions stated above, none of the tweedie models have density functions which have explicit analytic form or which can be written in closed form. however, the densities can be represented as infinite oscillating integrals, the methods of interpolation, inversion of the

cumulant generating function by the saddle point approximation method or evaluating the corresponding quasi likelihood of the distribution.

TABLE 3.2: *Quasi Likelihood Distributions for selected members of Exponential Families Densities*

Distribution	\mathcal{Q}
Normal	$-\frac{1}{2} \sum (y - \mu)^2$
Poisson	$\sum \{y \log \mu - \mu\}$
Binomial(κ)	$\sum \{y \log(\frac{\mu}{1-\mu}) + \log(1 - \mu)\}$
Gamma	$-\sum (\frac{y}{\mu} + \ln \mu)$
Inverse Normal	$\sum \{\frac{-y}{2\mu^2} + \frac{1}{\mu}\}$
Tweedie	$\frac{1}{\phi} [y \frac{\mu^{1-p}}{1-p} - \frac{\mu^{2-p}}{2-p}]$ for $p \neq (1,2)$

Tweedie with $p > 1$ have strictly positive means with $p > 2$ being continuous for positive Y and shape similar to the gamma, but more right skewed. Distributions with $p > 0$ are continuous on the real axis. for $1 < p < 2$ the distribution are supported on non negative real numbers and the distributions are mixtures of the poisson and gamma distributions with a discrete mass at zero. Tweedies are desired due to their ability to model both discrete and continuous data.

It is known that

$$\kappa''(\theta) = \frac{d\mu}{d\theta} = \mu^p$$

and mean by

$$\mu = \kappa'(\theta)$$

Such that

$$\mu^p = \frac{\partial^2 \kappa}{\partial \theta^2} \left(\frac{\partial \kappa}{\partial \theta} \right) = \frac{\partial \mu}{\partial \theta} \quad (3.15)$$

Taking reciprocals on both sides and intergrating with respect to μ gives

$$\theta = \begin{cases} \frac{\mu^{1-p}}{1-p} & p \neq 1, \\ \log \mu & p = 1 \end{cases} \quad (3.16)$$

by setting the arbitrary constant of integration to 0, and $\mu = \kappa'(\theta)$ gives

$$\kappa(\theta) = \begin{cases} \frac{\mu^{2-p}}{2-p}, & p \neq 2 \\ \log \mu, & p = 2 \end{cases} \quad (3.17)$$

The Tweedie densities can finally be written as

$$f_p(y; \mu, \phi) = a_p(y, \phi) \exp \left\{ \frac{1}{\phi} \left[y \frac{\mu^{1-p}}{1-p} - \frac{\mu^{2-p}}{2-p} \right] \right\} \text{ for } p \neq (1, 2) \quad (3.18)$$

A quasi likelihood is used by researchers if they don't know the density of the distribution but they know the mean and the variance.

For an observation, Q then

$$Q(y; \mu) = \int \frac{(y - \mu)}{V(\mu)} d\mu \quad (3.19)$$

[Wedderburn, 1974] showed that to fit a quasi likelihood function, only the mean and the variance relationship needs to be specified.

Following 3.19 then the Tweedie distribution has the following likelihood distribution when setting an arbitrary constant of integration to 0.

$$Q(\mu; y) = \int \frac{(y - \mu)}{V(\mu)} d\mu \quad (3.20)$$

$$= \int \frac{(y - \mu)}{\mu^p} d\mu \quad (3.21)$$

$$= \int \frac{y}{\mu^p} - \mu^{1-p} d\mu \quad (3.22)$$

$$= \int (y\mu^{-p} - \mu^{1-p}) d\mu \quad (3.23)$$

$$= \frac{y\mu^{1-p}}{1-p} - \frac{\mu^{2-p}}{2-p} \quad (3.24)$$

There is similarity of the last equation and 3.18 are similar only that we don't need to estimate the $a(y, \phi)$ that doesn't have closed form

[Dunn and Smyth, 2005] showed that the probability density function of the tweedie can be represented as

$$\log f_p(y; \mu, \phi) = \begin{cases} -\lambda, & \text{for } y = 0 \\ -\frac{-y}{\Upsilon - \lambda - \log y + \log W(y, \phi, p)}, & \text{for } y > 0 \end{cases} \quad (3.25)$$

where $\Upsilon = \phi(p-1)\mu^{p-1}$, $\lambda = \frac{\mu^{2-p}}{\phi(2-p)}$, and W is an example identified by as wrights generalised bessel function [Wright, 1933] and can be written as

$$W(y, \phi, p) = \sum_{j=1}^{\infty} \frac{y^{-j\alpha} (p-1)^{\alpha j}}{\phi^{j(1-\alpha)(2-p)^j \Gamma(-j\alpha)} \Gamma(-j\alpha)} \quad (3.26)$$

Where

$$\alpha = \frac{(2 - p)}{1 - p}$$

with the mean of the poisson gamma given as μ and its variance by

$$\text{Var}[y] = \phi\mu^p$$

Following [Dunn and Smyth, 2005] the probability of travelling zero distance is given by

$$\Pr(Y = 0) = \exp(-\lambda) = \exp\left[-\frac{\mu^{2-p}}{\phi(2-p)}\right] \quad (3.27)$$

3.4.5 Approximating tweedie densities using saddle point approximation

Various methods can be used to estimate a tweedie density including saddle-point, inversion or interpolation [Dunn, 2017, Dunn and Smyth, 2005]. In this thesis, we will consider the saddle-point approximation under GLM to approximate the starting values for the GEE.

There is part of the density that cannot be expressed in closed form, $b_p(y, \mu)$, as seen in equation 3.10, but can be replaced by a simple analytic expression such that

$$p(y | \mu, \phi) = \frac{1}{\sqrt{2\pi\phi y^p}} \exp\left\{\frac{-d(y, \mu)}{2\phi}\right\} \{1 + \omega(\phi)\} \quad (3.28)$$

as $\phi \rightarrow 0$ for the tweedie densities. The ratio is expressed as

$$\varsigma = b_p(y, \phi) \sqrt{2\pi\phi y^p} \quad (3.29)$$

such that

$$f_p(y | \mu, \phi) = \frac{1}{y} b_p(1, \iota) \exp \left\{ \frac{-d(y, \mu)}{2\phi} \right\} \quad (3.30)$$

where $\iota = \phi^{p-2}$ such that the ratio of the density to the saddlepoint is expressed as

$$\varsigma = b_p(1, \iota) \sqrt{2\pi\iota} \quad (3.31)$$

Which shows that ς is a function of p and not of μ and is a function of y and ϕ through ι .

Using Chebychevs interpolation method [Salzer, 1969], we can then estimate for any values of the parameter. The error is given by

$$f(x) - P_n(x) = \prod_{i=0}^n (x - x_i) \frac{f^{(n+1)}(\varpi(x))}{(n+1)!} \quad (3.32)$$

Such that we reduce the interpolation error by choosing x_i 's to minimize

$$||w(x)|| = \max_{x \in [a,b]} \left| \prod_{i=0}^n (x - x_i) \right| \quad (3.33)$$

3.4.6 Estimating the Parameters

We need to estimate β 's in order to fit a model. Under the GLM framework, the maximum likelihood are used to estimate the parameters using the Iterative weighted least square

as proposed by [McCullagh, 1984]. The likelihood function is usually defined by

$$L(\zeta | y) = \prod_{i=1}^n f(y; | \zeta) \quad (3.34)$$

where n is the sample size of the datasets, and ζ is the parameter of interest. The log likelihood is now defined as

$$\ell(\zeta | y) = \log L(\zeta | y) \quad (3.35)$$

$$\log \prod_{i=1}^n f(y | \zeta) \quad (3.36)$$

$$\sum_{i=1}^n \log f(y | \zeta) \quad (3.37)$$

Let a random variable with notation $Y \sim ED(\mu, \phi)$ come from the EDM, with dispersion parameter ϕ and location parameter μ . Then the log likelihood following equation 3.37 can be expressed as 3.38.

$$\mathcal{L}(\theta, \Phi | Y_1, \dots, Y_n) = \sum_{i=1}^n \left\{ \frac{Y_{it}\theta_{it} - \kappa(\theta_{it})}{a(\Phi)} + c(Y_{it}, \Phi) \right\} \quad (3.38)$$

To note however, the GEE are build from the GLM. This is through membership to the exponential family as described below. From a series of several equations and the chain rule, we differentiate the log-likelihood

for $1, \dots, n$ with respect to β values through a chain of equations expressed as

$$\frac{\partial L}{\partial \beta} = \left(\frac{\partial \ell}{\partial \theta} \right) \left(\frac{\partial \theta}{\partial \mu} \right) \left(\frac{\partial \mu}{\partial \eta} \right) \left(\frac{\partial \eta}{\partial \beta} \right) \quad (3.39)$$

We can easily show the derivation of the individual components by the following

Using equation 3.39 on equation 3.38, for the first component we can show that

$$\frac{\partial \ell}{\partial \theta} = \sum_{i=1}^n \frac{1}{a(\Phi)} [y_i - \mu_i] \quad (3.40)$$

Since the last component $c(Y_{it}, \Phi)$ differentiated with respect to θ is zero and $\kappa(\theta)$ w.r.t θ is $\kappa'(\theta)$. From equation 3.11 it is clear that $\kappa'(\theta)$ is replaced with μ .

The second component is expressed as

$$\frac{\partial \theta}{\partial \mu} = \frac{1}{V(\mu)} \quad (3.41)$$

This follows twice differential of $\mu = \kappa'(\theta)$ such that it equals to $\mu = \kappa''(\theta)$. Following 3.12 then this is expressed as $V(\mu_i)$ such that when you invert, you get equation 3.41

The third component is estimated as a link function expressed using equation 3.14 such that differentiating η_i w.r.t μ_i you get $g'(\mu_i)$ and inverting this leads to

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{g'(\mu_i)} \quad (3.42)$$

The last part uses this principle. let $\eta_i = \beta_0 x_{i,0} + \beta_1 x_{i,1} + \dots, \beta_j x_{i,j} + \dots, + \beta_p x_{i,r}$ where r is the rank of β . The derivative of η_i w.r.t β_j is given as x_{ij} .

A score equation from this derivation can now be expressed as

$$\frac{\partial \ell}{\partial \beta_j} = \frac{1}{a(\Phi)} \sum_{i=1}^n \frac{(y_i - \mu_i)}{V(\mu_i)} \frac{x_{ij}}{g'(\mu_i)} \quad (3.43)$$

To estimate the MLE we set the score equation to 0, and the $a(\Phi)$ doesn't need to be known.

For the estimating equations as defined by [Liang and Zeger, 1986] for a population average for GLM, the quasi-likelihood is given by

$$\Psi(\beta) = \sum_{i=1}^n \left\{ \frac{1}{v(\mu_i)} \frac{y_i - \mu_i}{a(\phi)} x'_{ij} D \right\} = 0 \quad (3.44)$$

whereby the first part of the equation is a generalization of the estimating equations of a GLM. The variance $V(\mu_i)$ can be decomposed into

$$V(\mu_i) = D(V(\mu_{it}))^{1/2} I_{(n_i \times n_i)} D(V(\mu_{it}))^{1/2} \quad (3.45)$$

We replace the identity matrix with a more general correlation matrix, say $R_i(\alpha)$, since the variance matrix for correlated data does not have a closed form. [Wedderburn, 1974] showed that for any choice of V_i , the estimate of β (say $\hat{\beta}$) is asymptotically normal and consistent such that mis-specification of the variance is not an issue in parameter estimation.

By modifying the Newton–Raphson algorithm using the Fisher scoring criteria, we can estimate the β 's. This procedure replaces the observed Hessian matrix with the expected

Hessian matrix. This is achieved by setting $R_i(\alpha)$ as an identity matrix and the scale parameter ϕ as estimated from the GLM.

To solve the estimating equations, we employ the iterative reweighted least squares algorithm, which is a modification of the Newton–Raphson algorithm such that the observed Hessian matrix replaces the expected Hessian matrix. The following approach is used to estimate β 's.

$$\hat{\beta}^{(r)} = \hat{\beta}^{(r-1)} - \left\{ \sum D_i^T v(\mu)_i^{-1} D_i \right\} \left\{ \sum D_i^T v(\mu)_i^{-1} S_i \right\} \quad (3.46)$$

$$D_i = D(V(\mu_{it})) D \left(\frac{\partial \mu_i}{\partial \eta} \right) X_i \quad (3.47)$$

$$S_i = y_i - g^{-1}(\hat{\eta}_i) \quad (3.48)$$

The iteration continues until the convergence set by the researcher is achieved.

3.4.7 Working correlation

Researchers have argued that there is always a true correlation that exists, however it is very hard to know or determine. Therefore, a working correlation matrix \mathbf{R} is produced to obtain an estimate of the covariance matrix. This correlation is of size $t \times t$ because they are measured at fixed time point during the survey. Each correlation matrix is of size $n_i \times n_i$.

A further assumption is that the correlation matrix \mathbf{R} depends on a vector of association parameters denoted by α therefore, the working correlation matrix can be defined by $R_i(\alpha)$ is now specified by the vector of unknown parameter α . Literature dictates that the choice of the correlation is fully the modelers preference and choice, its advisable to choose based on theoretical evidence.

However, we rely on the strength of correctly specifying of how the μ_i relates to the co-variates. And the covariance matrix converges to some fixed matrix in that the consistent parameter estimate is assured. Therefore loss of efficiency is reduced as we increase the clusters.

3.4.8 Independent Correlation Structure for Imbalanced Data

When it is suspect that no existing correlation in a dataset, then we can decide the correlation structure to be independent. This means we expect the same output as we could if we used the GLM approach. However, To fit a glm with a Tweedie distribution, the variance power can be estimated using the `tweedie.profile` command and link function are needed to be specified. The default link function is the canonical link function, with the logarithm link function also available.

The following command is used in `r` to fit a glm,

```
mod1 <- glm(dependent ~ independent, family=tweedie(var.power=p, link.power=0),
x=TRUE, data=data)
summary(mod1)
```

The above approach however doesn't compute the AIC, common in GLM to get the best fitting covariates that influences the dependent. The output from the summary only gives the p -values that shows the probability of rejecting the null under $\alpha= 0.05$. The output for the AIC is given as NA.

The implication herein is that you could have a set of many covariates which are significant but it is impossible to tell which set of combined covariates are best at predicting the dependent.

Therefore, for us to be able to draw additional scientific information that is relevant to our research, then the other approach would be to consider the independent correlation structure which means that we will still get the same results if we are using the glm, but better placed to get the best set of covariates for the model. Our approach will also enable us to pick the best set of covariates using the R -squared.

3.4.9 Models selection

The following set of 6 models were investigated in order to understand the effect of covariates on predicting outpatient healthcare spending in Kenya;

1. $\log \mu = \beta_0 + \beta_1 \text{age} + \beta_2 \text{wealthIndex} + \beta_3 \text{maritalStatus} + \beta_4 \text{education}$

2. $\log \mu = \beta_0 + \beta_1 \text{age} + \beta_2 \text{wealthIndex} + \beta_3 \text{education}$

3. $\log \mu = \beta_0 + \beta_1 \text{age} + \beta_2 \text{wealthIndex} + \beta_3 \text{maritalStatus} + \beta_4 \text{sex}$

4. $\log \mu = \beta_0 + \beta_1 \text{age} + \beta_2 \text{wealthIndex} + \beta_3 \text{maritalStatus} + \beta_4 \text{education} + \beta_5 \text{sex}$

$$5. \log \mu = \beta_0 + \beta_1 \text{age} + \beta_2 \text{wealthIndex}$$

$$6. \log \mu = \beta_0 + \beta_1 \text{wealthIndex}$$

Model 6 is a model on wealth index as a predictor for outpatient spending. The choice of modelling wealth index as a predictor was because it had the least individual QICu against the outpatient care spending, and is supported by literature [Awiti, 2014]. Model 5 is model controlling for age and wealth index. Age was also found to have individual low QICu compared to other covariates and therefore it was essence to find its effect with wealth index. Model 4 is model controlling for age, wealth index, marital status education and sex of the household head. Model 3 investigated control for age wealth index marital status and sex. Model 2 controlled for age wealth index and education. And finally model 1 controlled for age wealth index marital status and education.

In this model, since it was not possible to investigate all possible outpatient cost models, a systematic approach was adopted to find the most suitable model. First, a single predictor was developed and the QICu value examined for each. Models with lowest QICu were further examined Predictors were added successively in order of importance supported by the literature. Finally, through diagnostic check on the final models, we chose the one that fits the data adequately. There was no order in modelling the covariates. The model outputs are presented in table 4.14

In order to fit a tweedie GLM to the outpatient cost data, then it is appropriate to estimate the variance power. This is achieved through the profile log likelihood function with the MLE value corresponding to the most appropriate value of the variance function p with the respective 95% CI. Due to computational burden associated the method, a

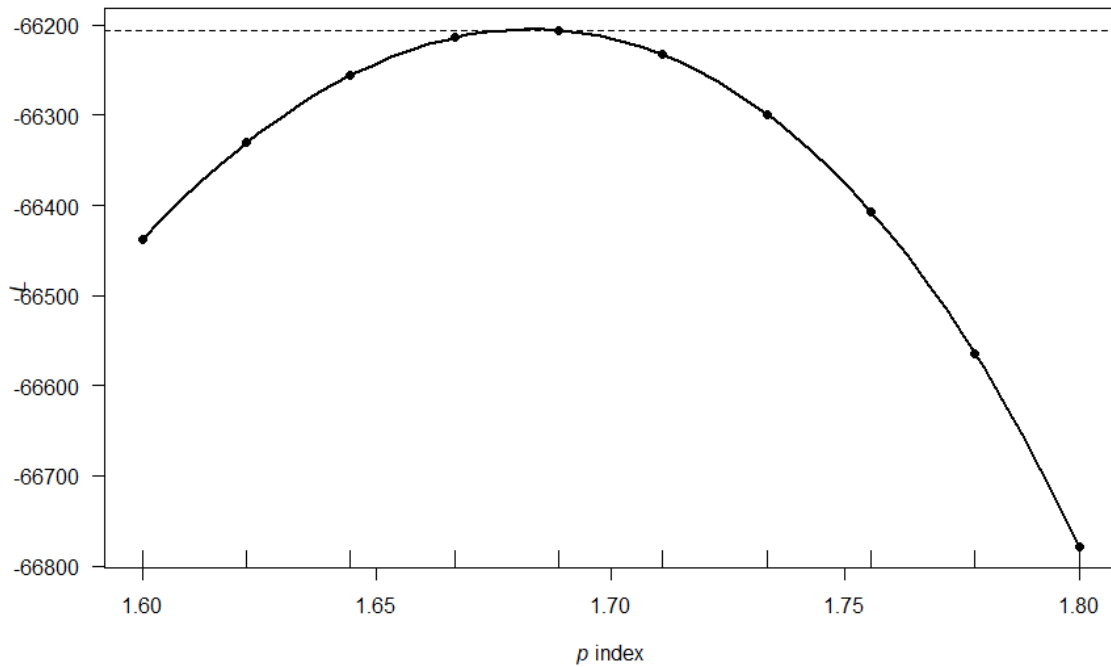


FIGURE 3.3: *The profile log-likelihood plot for cost of out patient care in Kenya using the model 1 covariates. The solid line is a saddle-point approximation of the P index from the data with a value of 1.68 and estimated 95% CI [1.67,1.69]*

cubic spline interpolation through these computed points is fitted. This is estimated as 1.68. Figure 3.3 shows the tweedie profile with the estimated index parameter and the confidence interval for the best fitting model.

3.5 Tweedie distributions in modeling clustered data using exchangable correlation structure and applications to distance-for inpatient care data

To model distance for inpatient care, we assume that it follows a gamma distribution.

Let R_i be distance recorded for a kenyan traveling seeking inpatient care.

We can assume that the distances traveled within a county during the survey period N follows a poisson distribution with mean λ such that if there is no distance covered, then $N = 0$ Finally Y represents total distance covered, and is represented as the poisson sum of the gamma random variables such that $Y = R_1 + \dots + R_N$ with the resulting distribution called the poisson-gamma distribution as defined by

3.5.1 Notations

Let \mathbf{y}_{it} be a vector of responses with a set of corresponding r covariates \mathbf{X}_{it} where i indexes K units of analysis $i = 1, 2, \dots, K$; and t indexes the time points $t = 1, 2, \dots, n_i$ for each unit. Thus the number of clusters observed is K . Also, $N = \sum n_i$ and is the total number of observations across all units. for inclusion of an intercept the first element of \mathbf{X}_{it} is set to 1.

Further, let $\mathbf{y}_i = [y_{i1}, y_{i2}, \dots, y_{in_i}]$ denote the corresponding column vector of observations on the responses variable for unit i and $\mathbf{X}_i = X_{i1}, X_{i2}, \dots, X_{in_i}$ indicate the $n_i \times r$ matrix covariates for unit i .

In our case application to model distance data, then we can define the following

- The response variable y_{it} is the distance travelled in a given county by any given individual seeking inpatient care
- With 47 data available for all the counties, then $K = 47$ and $i = 1, \dots, 47$
- The linear predictor $\eta_{it} = \mathbf{x}_{it}'\beta$

- define a link function used to relate the response to the linear combination of the covariates as

$$g(\mu_{it}) = \eta_{it}$$

- and the variance as a function of the mean, thus the distribution of the response variable is

$$\text{Var}[Y_{it}] = \phi V(\mu_{it})$$

- The correlation structure we are investigating is the exchangeable and the free specification which is an a hybrid of the unstructured

3.5.2 Exchangeable or Symmetrical Correlation for imbalanced data

The exchangeable or symmetrical correlation structure assumes that there is a common correlation within the observation in a given county. Thus all the correlation $R_i(\alpha)$ are all equal. Our data is based on cross-sectional data, thus they were collected at a specific time. This makes exchangeable a desired correlation to investigate under a tweedie distribution in GEE. For the exchangeable correlation structure, $R_i(\alpha)$ is defined as

$$R_{u,v} = \begin{cases} 1, & u = v \\ \alpha & \text{otherwise} \end{cases} \quad (3.49)$$

And is given in matrix form as

$$R_i = \begin{bmatrix} 1 & \alpha & \dots & \alpha \\ \alpha & 1 & & \alpha \\ \vdots & & \ddots & \vdots \\ \alpha & \alpha & \dots & 1 \end{bmatrix} \quad (3.50)$$

Following [Hardin, 2013] the exchangeable correlation structure uses the Pearson residuals

$$\hat{r}_{it} = \frac{(y_{it} - \hat{\mu}_{it})}{\sqrt{V(\hat{\mu}_{it})}} \quad (3.51)$$

from the current fit of the model to estimate the exchangeable correlation parameter. α is then estimated using the following

$$\hat{\alpha} = \frac{1}{\hat{\phi}} \sum_{i=1}^n \left\{ \frac{\sum_{u=1}^{n_i} \sum_{v=1}^{n_i} \hat{r}_{iu} \hat{r}_{iv} - \sum_{u=1}^{n_i} \hat{r}_{iu}^2}{n_i(n_i - 1)} \right\} \quad (3.52)$$

In this study, a Tweedie distribution is used to model predictors of distance for inpatient care. The justification to use the Tweedie distribution is provided in Fig 3.4, which shows that distance as the dependent variable has a discrete mass at zero and a continuous characteristic. In addition, Table 3.3 shows that the data are right-skewed with a skewness value of 4.80. We analyzed the distance with the covariates to determine the best combinations to explain any existing association.

An important property of Tweedie under GEE, is its ability to accommodate both correlation and right skewedness which is a characteristic of our continuous data. This approach is used in this paper and it complements the work of [Swan, 2006], who used the AR(1)

TABLE 3.3: *Descriptive analysis of Distance for inpatient care*

Statistic	Distance travelled (km)
Minimum	0
1st Quarter	3
Median	10
Mean	32.1
3rd Quarter	30
Maximum	700
Shape	Right Skewed
Skewness	4.8

correlation structure. Tweedie regression models allow relation of the mean of distance to the selected covariates. This allows the mean of distance to be modeled as a linear function of covariates using the log link given by

$$\log(\mu_i) = \beta_0 + \beta_1 x_i + \dots + \beta_n x_n \quad (3.53)$$

where β_j vectors are regression coefficients that corresponds to the x_j vectors of covariates, all fitted based on Tweedie EDM. Means are calculated to assess the relationship between covariates and the distance.

To fit the models, we need to estimate the index parameter p and the β 's from the Tweedie distribution using the GLM framework. This could be computationally difficult but the R package Tweedie [Dunn and Smyth, 2008] and statmod [Giner and Smyth, 2016] fit this easily. The calculated index parameter as calculated by the software is shown by Fig 2 in S2 Appendix. To estimate the β 's, we apply the approach of [McCullagh, 1984] called the Iterative re-weighted least square (IRLS) method in the GLM's.

These β 's are then updated under the GEE framework as follows.

Following [Hardin, 2013], quasi-likelihood estimating equations for GLM without any preference for mean and variance function with 47 clusters are expressed as

$$\Psi(\beta) = \sum_{i=1}^n \left\{ x'_{ik} T \frac{1}{V(\mu_i)} \frac{y_i - \mu_i}{a(\phi)} \right\} \quad (3.54)$$

for $k = 1, 2, \dots, 47$ where T is a diagonal matrix of derivatives

$$\frac{\partial \mu_i}{\partial \eta_i} \quad (3.55)$$

$V(\mu_i)$ is an $n_i \times n_i$ diagonal matrix which can be decomposed into

$$V(\mu_i) = T [V(\mu_i)^{1/2} I_{(n_i \times n_i)} V(\mu_i)^{1/2}] \quad (3.56)$$

The observations in the clusters are treated as independent, but our focus is on the population average for which both the mean and the variance are averaged over all the clusters. Following [Liang and Zeger, 1986], the identity matrix in the above is replaced with a more general correlation matrix since the variance matrix for data which is correlated doesn't have a diagonal form, as follows

$$V(\mu_i) = T [V(\mu_i)^{1/2}] CM_i(\alpha) T [V(\mu_i)^{1/2}] \quad (3.57)$$

where the correlation matrix $CM_i(\alpha)$ is estimated by vector α . Proper specification of the $CM_i(\alpha)$ then the $\hat{\beta}$ are consistent and asymptotically normal. To achieve more efficiency, then it is necessary to include a hypothesized correlation structure within the clusters. The structure that is suitable for our clustered data without any time

dependence is the exchangeable given by

$$CM_{ik}(\alpha) = \begin{cases} 1 & \text{if } i = k \\ \alpha & \text{if } i \neq k \end{cases} \quad (3.58)$$

From the above, equation 3.58 the matrix is given as

$$CM_i = \begin{bmatrix} 1 & 0.045 & \dots & 0.045 \\ 0.045 & 1 & & 0.045 \\ \vdots & & \ddots & \vdots \\ 0.045 & 0.045 & \dots & 1 \end{bmatrix} \quad (3.59)$$

For the multiple regression analysis, this work adjusted for household head gender, education, age, household size, and wealth index. It further adopted the exchangeable correlation structure under the GEE approach using the Tweedie distribution.

This work examined the regression model fit to select the best model using the quasi-likelihood based information criterion, or QICu, proposed by [Hardin, 2013], which is an extension of the QIC proposed by [Pan, 2001], following the Akaike information criterion developed for the GLM by [Akaike, 1998].

The QICu imposes a penalty based on model complexity to ensure that only a few covariates are used to achieve model parsimony. Data in .sav format were imported into R statistical software version 3.6.3 [R Core Team, 2017] for cleaning, reformatting, recoding, and analysis.

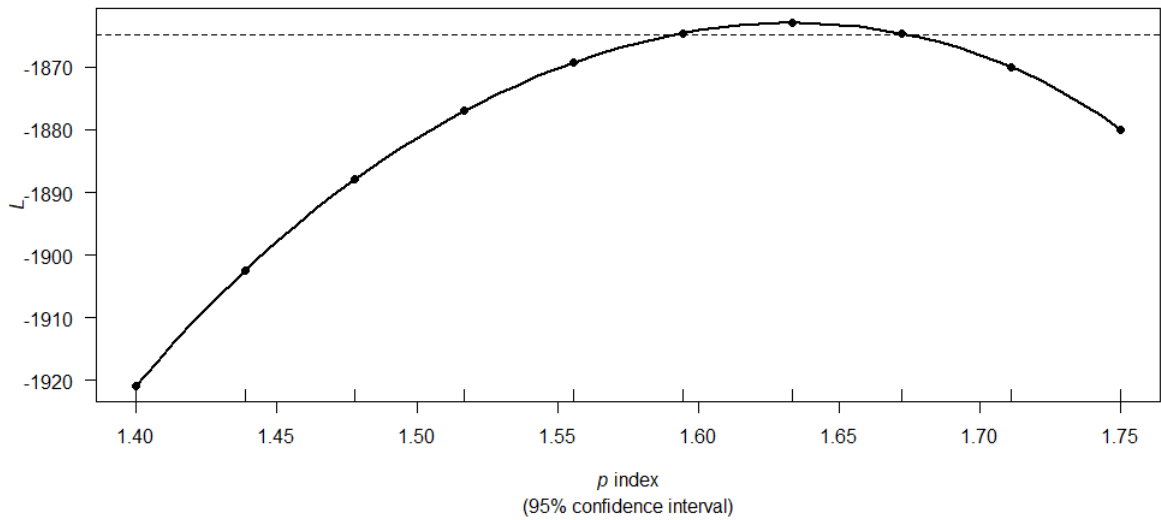


FIGURE 3.4: *The profile log-likelihood plots for the distance travelled for inpatient care in Kenya using gender of the household head, household size, and education as covariates. The plot estimated the p as 1.63 (1.59,1.67), with the dots representing 95%. The solid line is a cubic-spline smooth interpolation joining the points.*

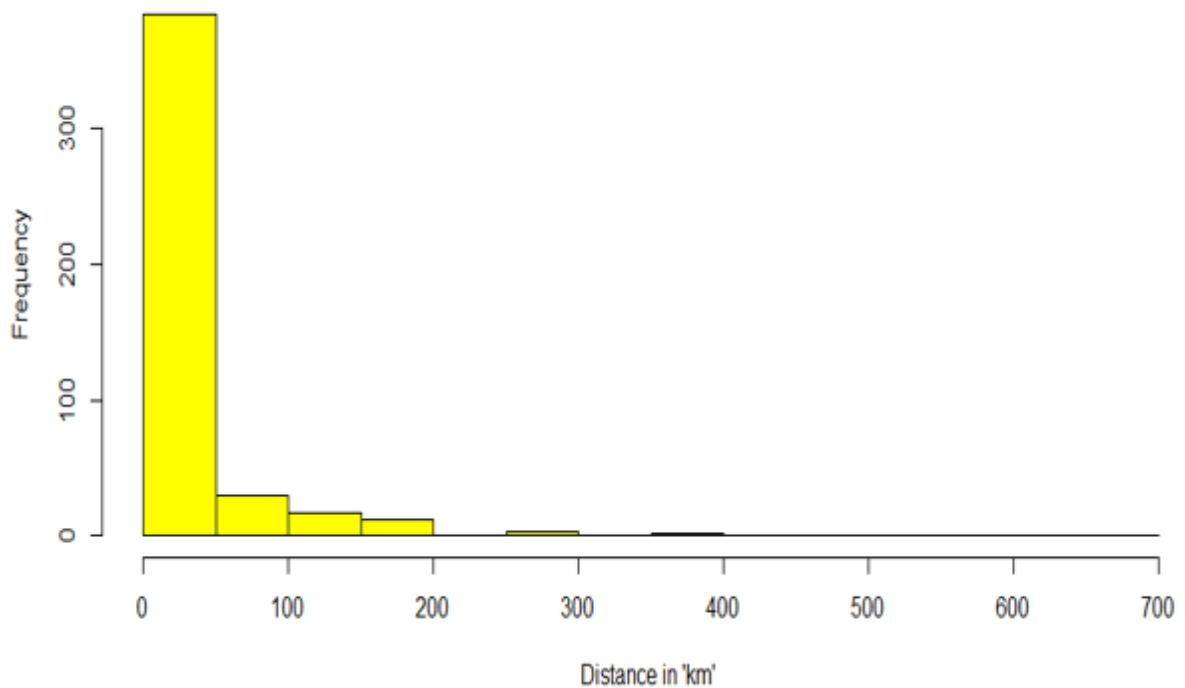


FIGURE 3.5: *Histogram of Distance travelled to seek inpatient care in Kenya*

Chapter 4

Results and Discussion

4.1 Introduction

This chapter presents the results

4.2 Effect of Bacterial Vaginosis (BV)-HIV-1 Co-Existence on Maternal and Infant health: A Secondary Data Analysis Results

Table 4.1 presents the Demographic characteristics of the women. The mean age was 23.5 year (range 17-39) for the BV-exposed infants and 24.1 years (15-38) for the unexposed infants. Among BV-exposed women 27.2% had not completed the 8 years of primary education, 19.6% had completed 8 years and 53.3% had > 8 years education compared

to 35.9%, 17.2% and 46.9% respectively among the unexposed women, however these differences were not significant ($p=0.18$). The majority of the women were married, 81% of the BV-exposed group and 73.4% of the unexposed group, and these differences were not significant ($p=0.18$).

Table 4.2 presents the selected maternal morbidity incidences in relation to BV exposure. There were significant differences between the two groups.

Among the women with data on maternal conditions at delivery, compared to unexposed women, women diagnosed with BV had a higher prevalence of maternal complications (18/169 [10.6%] vs. 7/179 [4%]). Women with BV had a higher rate of hospital admission (35/135 [26%]) compared to women without BV (21/179 [15%]). The mean duration of hospital admission was 1.3 ± 3.2 days among women with BV compared to 0.7 ± 2.7 days among women with BV. Exposed mothers were 2.93 times likelier (95% CI, 1.24–7.71) to report adverse maternal conditions and 1.95 times likelier (95% CI, 1.08–3.51) to be admitted to the hospital at birth ($P=0.02$).

The relationship between BV and log viral load is shown in Table 4.3. There was no significant difference between the two groups regarding viral load (OR, 1.21; 95% CI, 0.91–1.60, $P=0.192$).

Sample of distribution of the selected neonatal characteristics are depicted in Table 4.4. Neonates exposed to BV were comparable to unexposed babies in terms of gestational age, Apgar score, and anthropometric measures of weight and height. There were fewer male infants among babies exposed to BV compared to unexposed babies (81/179 [45%] vs. 99/178 [55%]). However, there were more deaths in the exposed group (9 [5%]) than

TABLE 4.1: Comparison of the Mothers Demographic and selected maternal Characteristics between the two Groups

Variable	Exposed	Unexposed	RR(95% CI)	P-value
<i>Maternal demographics</i>				
Age: Mean (Range)	23.5(17-39)	24.1(15-38)		0.21
Marital Status: N (%)				
Single	28(15.2%)	42(21.9%)		0.18
Divorced/Separated/Widowed	7(3.8%)	9(4.7%)		
Married	149(81.0%)	141(73.4%)		
Education: N (%)				
< 8 Years	50 (27.2%)	69(35.9%)		0.18
8 Years Complete	36(19.6%)	33 (17.2%)		
>8 Years	98(53.3%)	90 (46.9%)		
Number of pregnancies: Mean (Range)	2(1-8)	2(1-6)		0.22
Number of live births: N (%)				
0	73(39.7%)	75(39.1%)		0.74
1 and 2	87(47.3%)	86(44.8%)		
>3	24(13.0%)	31(16.1%)		
Birth Outcome: N (%)				
Live	168 (97.7%)	181(98.4%)		0.31
Intra-partum Deaths	2 (1.2%)	3 (1.6%)		
Still Births	2 (1.2%)	0(0%)		
Any STD: N (%)				
Yes	52 (28.3%)	59 (30.7%)		0.71
No	132 (71.7%)	133 (69.3%)		
<i>Maternal Characteristics</i>				
Log viral load per milliliters:Mean(SD)	10.64(1.79)	10.37(1.88)		0.19
Viral load> 39482: N (%)				
Yes	79(53.3%)	76(46.3%)	1.15(0.92-1.44)	0.25
No	69(46.7%)	88(53.7%)		
Admitted after birth?: N (%)				
Yes	32(19.3%)	16(9.7%)	1.99(1.14-3.48)	0.02
No	134(80.7%)	149(90.3%)		
Adverse maternal Conditions?: N (%)				
Yes	16(9.7%)	7(4%)	2.45(1.04-5.81)	0.04
No	149(90.3%)	170(96%)		

TABLE 4.2: Outcomes of maternal morbidity incidences

	Bacterial Vaginosis Status		OR (95% CI)	P-value
	Present	Absent		
Adverse maternal conditions, % (n)	10.6 (18/169)	4 (7/179)	2.93 (1.19–7.20)	0.02
Maternal hospital admissions, % (n)	26 (35/135)	15 (21/179)	1.95 (1.08–3.51)	0.02
Mean duration of admission, days	1.3±3.2	0.7±2.7		

OR comparison are made to the bacterial vaginosis unexposed. Data are reported as proportions (of patients with valid data) or mean \pm standard deviation. P-values were derived from Fisher's exact test. Maternal admissions were the ones immediately after birth

TABLE 4.3: OR of bacterial vaginosis and Log viral load

	OR (95% CI)	P-value
Viral load	1.21 (0.91–1.60)	0.19

in the unexposed group (5 [2.8%]). On average, BV unexposed infants had a higher mean body length (48.6±2.5 cm) than BV exposed infants (48.1±4.3 cm). Overall, 5 (3%) of 169 babies exposed to BV had an LBW (<2500 g) compared to 1 (1%) of 178 unexposed infants.

TABLE 4.4: Distribution of neonatal characteristics

	Bacterial vaginosis status	
	Exposed	Non-exposed
Died, n (%)	9 (5.4)	5 (2.8)
Birth weight per 100 g	31±5.5	32±5.0
Length, cm	48.1±4.3	48.6±2.45
Gender (male), n (%)	81 (46.5)	99 (55.6)
Head circumference, cm	35.2±1.49	35.2±1.59
Maturity, weeks of gestation	39.5±2.35	39.8±2.03
Apgar score	9.63±1.29	9.71±0.87
Maturity, Dubowitz score	57.7±8.2	57.8±8.11
Low birth weight (<2500 g), n (%)	5 (3)	1 (1)

Table 4.5 presents the association between exposure to BV and birthweight which had

been shown in other studies to be the most important independent predictor of BV, whereby a non-linear association was reported. Exposed neonates had 0.96 times odds of weight compared to unexposed. However, the results not significant at the 0.05 level.

TABLE 4.5: *Bacterial vaginosis and birth weight as a continuous variable*

	OR (95% CI)	P-value
Birth weight	0.96 (0.92–1.00)	0.08

Evaluation of any morbidity incidences among the neonates is presented in Table 4.6. It is not surprising that no morbidity incidence are associated with BV. This is because babies interact with their mothers biome during birth and therefore the effects of BV are evident after birth.

TABLE 4.6: *ORs of morbidity incidence among the neonates*

	OR (95% CI)	P-value
Jaundice	1.2 (0.58–2.52)	0.62
Conjunctivitis	1.39 (0.57–3.50)	0.47
Lymphadenopathy	0.70 (0.30–1.57)	0.39
Respiratory distress	1.03 (0.12–8.66)	0.98
Skin rash	0.93 (0.48–1.77)	0.82
Prematurity	0.70 (0.09–4.29)	0.7
Asphyxia	1.06 (0.25–4.55)	0.93
Pneumonia	2.13 (0.20–46.12)	0.62
Sepsis/meningitis	0.63 (0.13–2.60)	0.53
Other abnormality on exam	1.29 (0.34–5.28)	0.71

To assess overall morbidity of incidences with the percentage on, ever reported, among infants during the year, the study results are presented in Table 4.7. In the Chi-Square(χ^2) test of association between BV and the various morbidities, only hepatomegaly and having a cold showed statistical significance. No other morbidities assessed exhibited any association at ($P < 0.05$).

TABLE 4.7: *Overall morbidity incidences reported and correlation analysis of bacterial vaginosis and morbidity incidences*

	Overall morbidity incidence, n(%)	Bacterial vaginosis status		Chi-squared test
		Exposed, n (%)	Non-exposed, n (%)	P-value
Pneumonia	111 (34)	47 (30)	64 (37)	0.2
Ear infection	32 (10)	17 (10)	15 (27)	0.5
Blood in stool	38 (12)	21 (13)	17 (10)	0.3
Lymphadenopathy	140 (43)	62 (39)	78 (46)	0.3
Encephalopathy	3 (1)	1 (1)	2 (1)	0.6
Sepsis	22 (7)	11 (7)	11 (6)	0.8
Conjunctivitis	78 (24)	43 (27)	35 (21)	0.1
Dehydration	5 (2)	2 (2)	3 (1)	0.7
Wheezing	70 (21)	30 (19)	40 (23)	0.3
Hepatomegaly	47 (14)	14 (9)	33 (19)	0.007
Cold	283 (86)	129 (82)	154 (90)	0.04
Otitis	21 (6)	9 (6)	12 (7)	0.6
Fever	243 (74)	111 (70)	132 (77)	0.2
Cough	290 (88)	134 (84)	156 (91)	0.1
Diarrhea	20 (6)	7 (4)	13 (8)	0.2
Thrush	96 (29)	43 (27)	53 (31)	0.5
Vomiting	146 (45)	72 (46)	74 (43)	0.6
Difficulty feeding	146 (45)	72 (46)	74 (43)	1.6
Heat rash	120 (37)	58 (37)	62 (36)	0.9
Fungal rash	64 (20)	33 (21)	31 (18)	0.5
Eczema/dermatitis	62 (19)	28 (18)	34 (20)	0.6
Scabies	63 (19)	30 (19)	33 (19)	0.99
Mouth ulcers	15 (5)	8 (5)	7 (4)	0.7

Results of multiple logistic regression analysis are described in Table 4.8. The study findings shows that at 6 months, infants of BV exposed mothers were 3.08 times likelier to have bloody stool and 2.94 times of being dehydrated. They were also more likely to vomit 1.64 times and had higher odds of mouth ulcers at 12.8 times. At 12 months, exposed infants were 1.81 times likely to be dehydrated and were more likely to vomit with odds of 1.39. Our results depicts a trend of decrease in morbidity with growth of an infant.

Additionally, there were higher reports of hospitalization (OR 1.12 95% CI(0.61,1.68), p=0.96) among the exposed though results not statistically significant, and higher clinic visits among the exposed (OR 1.26, 95% CI(1.01,1.61),p=0.07) though results are not statistically significant.

TABLE 4.8: Predictors of bacterial vaginosis at 6 and 12 months with corresponding 95% Confidence Intervals (CI) and *p*-values

		Six months		Twelve months	
		OR (95% CI)	P-value	OR (95% CI)	P-value
Respiratory infections	Pneumonia	1.14 (0.75–1.72)	0.54	1.20 (0.88–1.65)	0.25
	Ear infection	0.55 (0.18–1.54)	0.27	0.63 (0.33–1.17)	0.14
	Wheezing	0.67 (0.34–1.27)	0.22	0.87 (0.59–1.27)	0.47
	Cold	1.09 (0.88–1.35)	0.44	0.99 (0.84–1.16)	0.88
	Otitis	1.10 (0.60–2.01)	0.76	1.10 (0.79–1.55)	0.57
	Cough	0.97 (0.77–1.23)	0.81	0.93 (0.78–1.10)	0.39
Gastrointestinal infections	Stool with blood	3.08 (1.11–10.00)	0.04	1.19 (0.67–2.12)	0.56
	Dehydration	2.94 (1.44–6.37)	0.01	1.81 (1.05–3.19)	0.03
	Diarrhea	0.72 (0.43–1.22)	0.23	0.64 (0.45–0.89)	0.01
	Vomiting	1.64 (1.06–2.56)	0.03	1.39 (1.01–1.92)	0.04
	Mouth ulcers	12.8 (2.27–241.21)	0.02	2.34 (1.00–6.03)	0.06
Other infections	Lymphadenopathy	0.74 (0.55–0.99)	0.04	0.65 (0.52–0.82)	0.01
	Encephalopathy	0.55 (0.07–3.57)	0.53	0.51 (0.07–2.74)	0.45
	Sepsis	1.27 (0.55–2.96)	0.57	1.18 (0.53–2.65)	0.68
	Conjunctivitis	1.32 (0.84–2.08)	0.24	1.41 (0.95–2.10)	0.09
	Difficulty feeding	0.74 (0.49–1.12)	0.16	0.78 (0.63–0.97)	0.02
	Heat rash	0.79 (0.55–1.13)	0.2	0.75 (0.56–1.01)	0.06
	Fungal rash	1.04 (0.62–1.77)	0.87	1.15 (0.75–1.78)	0.51
	Eczema/dermatitis	0.95 (0.76–1.19)	0.67	0.87 (0.73–1.04)	0.13
	Scabies	0.78 (0.37–1.62)	0.5	1.04 (0.69–1.55)	0.86
	Hepatomegaly	0.47 (0.19–1.05)	0.08	0.56 (0.32–0.94)	0.03
Fever	0.75 (0.57–0.99)	0.04	0.89 (0.73–1.07)	0.22	
Thrush	0.94 (0.64–1.37)	0.74	0.92 (0.67–1.28)	0.63	

4.2.1 Mortality in the first twelve months of life

We compared survival between infants whose mothers were exposed and those whose mothers were not (Figure 4.1). There was no significant difference in the mortality distribution between the two groups ($p=0.65$); however, the graph showed a trend of higher mortalities in the BV exposed group.

4.2.2 Discussion

The results of this study suggest that exposure to BV has a significant effect on the incidence of morbidity among babies and their mothers. It was important to analyze the status of the baby after birth, because healthy babies grow well, 6 months because that

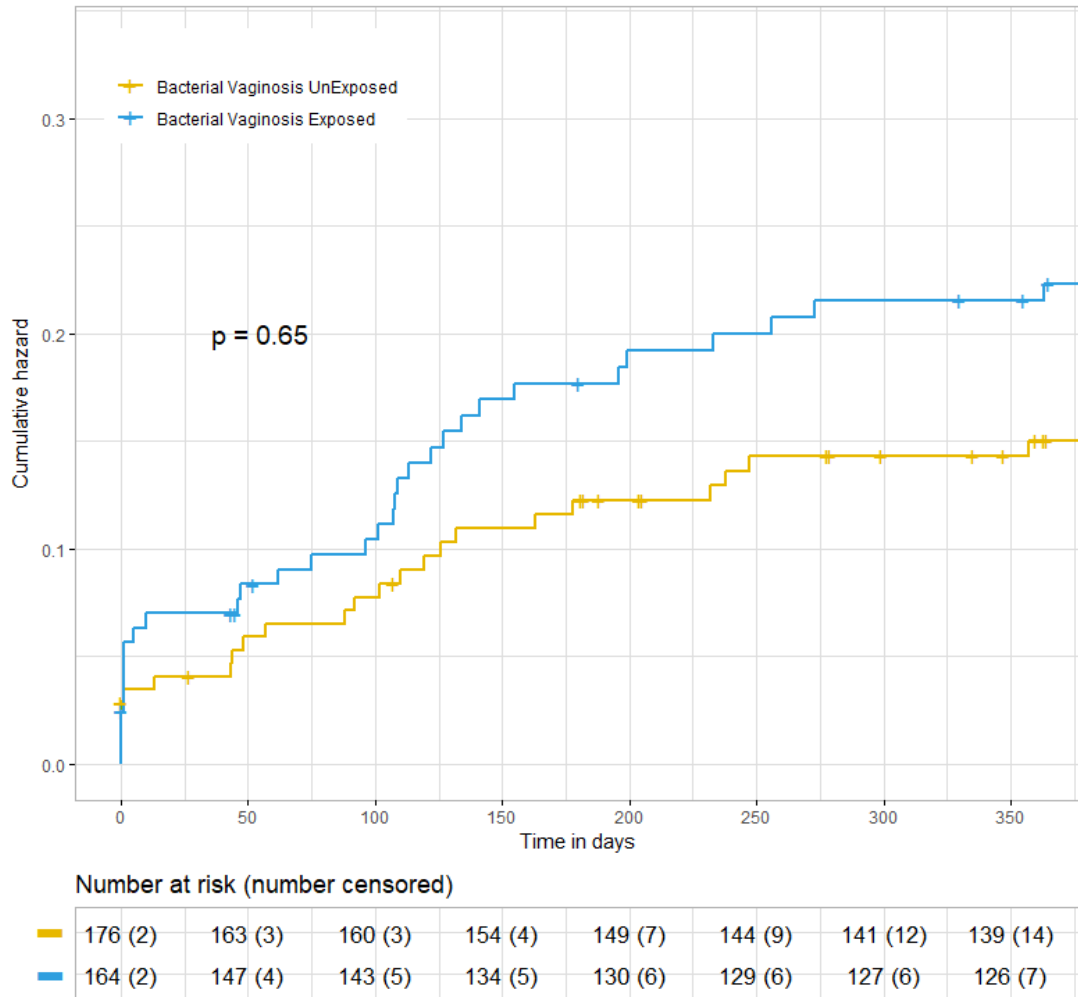


FIGURE 4.1: *Kaplan–Meier analysis of infant mortality over 12 months between infants with maternal exposure/ non-exposure to bacterial vaginosis*

is the time a baby is initiated into complementary feeding and 1 year as it marks end of infancy and beginning of childhood. As supported by the literature, in the assessment of neonates, no morbidity showed statistically significant results even after adjustment for other morbidities. This has previously been reported by [Nduati et al., 2000]. However, the findings of the study showed adverse neonatal effects among the exposed subjects, similar to that reported by [Dingens et al., 2016]. Several studies have established an association between BV and various morbidities. For example, [Hillier et al., 2018] reported that exposed women had twice the chance of giving birth to children with LBW. Our analysis (which yielded non-significant results at the 0.05 alpha level) showed a 0.95

times likelihood of LBW among exposed subjects. The lower mean birth weight in the exposed group is also a good pointer to the consistencies in association as reported in other related studies [[Freitas et al., 2017](#), [Isik et al., 2016](#), [Thorsen et al., 2006](#)]).

Hospital visits (including admissions and non-admissions) for different morbidities have been investigated by several authors. Maternal hospitalization after birth has drawn the interest of researchers and particularly in developed countries, there has been a trend toward shortening the postpartum stay in hospital. Some of the factors include cost and availability of hospital beds for other mothers in need of care. [[Evans et al., 2008](#)] investigated the impact of early discharge on outcomes among infants and found no differences in the outcomes of early and late discharge. However, the authors conducted a meta-analysis on the same data and found no study that reported any differences, even though there was an international trend toward shortening the postpartum length of stay in hospitals among women who have undergone vaginal delivery to improve the mother's sleep, for proper bonding of the mother and infant, and to protect the infant-mother dyad from nosocomial infections [[Benahmed et al., 2017](#)].

Our analysis showed a 1.95-fold increase in the frequency of maternal hospital admissions among exposed subjects compared to non-exposed subjects. Although no benefit or risk has been associated with a longer stay in hospital, long admissions always have a cost implication. The hospitalization of infants has been investigated previously. [[Jones et al., 2018](#)] investigated hospitalization as a result of general specific causes in Europe and reported higher odds among infants with jaundice and difficulty feeding. While this study does not specify the causes of hospitalization, it compares the odds of hospitalization

between the two groups. From the study results, there were no differences in hospitalization at 6 months; however, higher odds among exposed subjects for both admissions and non-admissions were reported.

Although not statistically significant at the 0.05 alpha level, the results show a direction and strength of the effect, with a 1.12-fold increase among exposed and hospitalized subjects. Hospitalization was a very key indicator of health outcomes because in as much as an infant could be hospitalized due to specific morbidities, he/she may end up being diagnosed with a different morbidity which will also be treated. This could result in what we refer to as reverse causality, in which an unexposed subject could show higher odds for a disease than an exposed subject. This is evidenced by morbidity incidences regarding hepatomegaly, diarrhea, and difficulty feeding which showed higher odds among unexposed subjects at 12 months. Diarrhea, in contrast to our findings, has been found to be a good predictor of infant morbidity in other studies [[Anigilaje, 2018](#), [Berger et al., 2007](#), [Chang et al., 2018](#), [Khalil et al., 2019](#)].

Lymphadenopathy and fever at 6 months, showed higher odds among unexposed subjects; as this result was incongruent to that reported in the literature, we performed further analyses. Infant hospitalization does not only have a negative effect on the development and physical growth of an infant, but also results in psychological distress and loss of parenting on the part of the mother [[Lean et al., 2018](#)]. This is one of the indicators of childhood morbidity as infants with longer periods of hospitalization tend to show higher morbidity rates due to the risk of disease exposure at care facilities, particularly in developing countries [[Shiva et al., 2017](#)].

At 6 months, there was a 3.08-fold increase in the passage of bloody stool among exposed subjects, the results of which were significant at the 0.05 alpha level. A growing body of evidence has linked this with infant colitis and intestinal infections [Murphy, 2008], which were not assessed as morbidities in the present study. Though not direct link with BV, our results suggest a causal relationship.

Another morbidity of interest is dehydration which yielded significant results at both 6 and 12 months. [Finberg, 2002] defined dehydration in infants as a loss of water and salt or extracellular fluid, caused by bacterial and viral agents. Our results showed a 2.94-fold and 1.181-fold increase in the rate of dehydration among exposed subjects at 6 and 12 months, respectively. Vomiting has been associated with a 1.64-times and 1.39-times increase in odds among exposed subjects at 6 and 12 months, respectively, the results of which were significant at the 0.05 alpha level. Some authors have associated this with a lot of infant discomfort, thus hindering their optimal proper growth [Gibson, 1959].

Finally, the other morbidity of interest is mouth ulcers. These, which vary in size, are open wounds that spread across the mouth lining of an infant and have diverse effects on their growth. Some direct effects include difficulty feeding as a result of pain, burning, and irritation of the mouth. A 12.8-fold increase at 6 months and a 2.34-fold increase at 12 months among exposed subjects demonstrates the seriousness of the effects of BV. This is a morbidity that requires proper intervention to enable proper growth of the infant.

The maternal viral load was investigated, and 1.21-fold odds were reported among exposed subjects, the results of which were not statistically significant. Our results were consistent with those reported by [Burns et al., 1997] who reported an association and a 3-fold odds of vaginal candidiasis among women infected with HIV but with low CD4 counts. In

addition, a statistically significant association was reported by [Atashili et al., 2008] between HIV and BV.

Although there is no direct link between the CD4 count and viral load, low levels of the latter are desirable. [Jamieson et al., 2001] reported on severe BV among women infected with HIV. The viral load is a very essential component, particularly in the context of HIV and an undetectable viral load is very desired as it reduces the risk of transmission, especially from mother to child.

[Mbori-Ngacha et al., 2001] reported that any BV could be treated during pregnancy ; however, studies have shown that treatment does not scale down the adverse effects associated with preterm birth and neonatal risks [Brocklehurst et al., 2013, Carey et al., 2000].

There has not been any published literature linking mortality directly to BV. However, there is an indirect link through the risk factors of BV. Preterm birth, pregnancy complications, and prematurity are risk factors that greatly affect neonates and infant mortality as reported in the Demographic and Health Surveys [Leidman et al., 2018] and by other authors [Chaim et al., 1997, Ukah et al., 2020].

While, in their previous study, [Nduati et al., 2000] excluded intrapartum deaths, stillbirths, and abortions, our analysis captured intrapartum deaths. The Kaplan–Meier analysis through the cumulative hazard plot showed no differences in the hazard of mortality ($p=0.65$); however, the graph showed a trend toward a higher mortality rate in the BV exposed group. Therefore, this means that the risk of mortality still exists among infants whose mothers are exposed to BV.

In conclusion, the study results were consistent with those reported in the literature and added further knowledge in the area of HIV. This study showed that BV among HIV-exposed women could result in infants who are more vulnerable to several infections due to a compromised immune status. Assessing the risk of BV infection in HIV-positive women could be a step in the right direction of developing policies targeting limited resource countries that could finally mitigate the fatal adverse outcomes on mothers and their infants.

4.3 Skewed Logit Model for Correlated data and its application to infant morbidity results

The preliminary analysis showed that 148 (45%) infants were born to women who tested positive for BV while the remaining 179 (55%) were born to women who tested negative for BV. 185 (57%) infants were breastfed, while 142 (43%) were formula-fed. 168 (51%) were males and 159 (49%) were female. 61 (19%) of the infants were HIV-positive, while 266 (81%) were HIV-negative.

It was of scientific interest to model the effects of BV on the marginal probability of an infant suffering from different morbidities in the first six months of life. Assuming morbidity incidence as the response, our data had the number of morbidity incidences recorded in a given month for each infant. Zero was recorded if no incidences occurred.

The study sought to assess whether children born to women who tested positive for BV were more likely to have a higher morbidity incidence than their counterparts and if the

effects would change with time. The literature has shown that BV has more effects during the first months after birth as the child continues to build immunity as they grow. Also, we expect children who gain weight consistently to have fewer morbidity incidences than those babies who take time to gain weight.

TABLE 4.9: *BV with morbidity incidences reported from month one to six for both BV-exposed and unexposed babies in the Nairobi data survey*

Time in Months	BV present(n=148)	BV absent(n=179)	Total(n=327)
1	115(78%)	85(47%)	200(61%)
2	97(66%)	84(47%)	181(55%)
3	97(66%)	85(48%)	182(56%)
4	92(62%)	101(56%)	193(59%)
5	86(58%)	96(53%)	182(56%)
6	79(53%)	103(58%)	182(56%)

The frequency of morbidity incidence seemed to decrease evenly in the BV present group. This was not the same in the BV absent group, which evidenced increases and decreases in morbidities in the different months considered(Table 4.9). It was, therefore, important to examine the effect of BV on infant morbidity over time. In order to correctly estimate the marginal effect of the parameters of interest to be estimated, a distribution had to be chosen for the dependent variable, which did not involve assuming a specific distribution would apply. Thus, we considered the following logistic model:

$$\begin{aligned} \text{logit}(\pi_{ik}) = \log\left(\frac{\pi_{ik}}{1 - \pi_{ik}}\right) = & \beta_0 + \beta_1\text{BV}_i + \beta_2\text{HIV}_i + \beta_3\text{feeding}_i + \beta_4\text{male}_i \\ & + \beta_5\text{time}_k + \beta_6\text{weight}_i + \beta_{51}\text{time}_k \times \text{BV}_i \end{aligned} \quad (4.1)$$

for $k = 1, \dots, 6$, $i = 1, \dots, 327$ where $\text{Time}_k = 1$ if the reference moth is under consideration and zero otherwise. $\text{Male}_i = 1$ if the i^{th} child is male and 0 if female, $\text{HIV}_i = 1$ if the

child tests positive to HIV and 0 otherwise, $\text{Breastfed}_i = 1$ if the child was randomized to the breastfeeding group and 0 if randomized to the formula feeding group, $\text{BV}_i = 1$ if the mother tested positive for BV and 0 if she tested negative, and Weight_i is recorded continuously for each infant across the six months. Data on morbidities from birth were included in the month 1 tally, and not as an independent time period, since morbidities due to BV on neonates was found to be insignificant in previous research [Mwenda et al., 2021b, Nduati et al., 2000]. Interaction of time and BV was used to assess changes in immunity over the time of exposure.

Our data can effectively account for the within-subject correlation. Hence, we consider the following correlations structures in terms of independence, as well as whether they are exchangeable, AR(1), M-dependent, and unstructured. This study was interested in comparing two models, the skewed logit-GEE and the standard GEE, when the response is assumed to be asymmetric.

We assessed different correlation structures and all our parameter estimates were within the acceptable standard error ranges. However, measurements which were not taken for the same individual exhibit lower correlation and follow a pattern imposed by AR(1), thus for the interpretation of our work, AR(1) was adopted. For the M-dependent variable, the default $m = 1$ was used.

Very few iterations are needed for the convergence of the models in GEE. Therefore, we initially set the maximum iterations to 50, however, the models with M-dependent variable and AR(1) correlation structure did not converge. We increased the maximum iterations to 200 and this achieved convergence. More precisely, independence converged

TABLE 4.10: *Regression Parameter Estimates with Model-Based and Empirical Standard Errors (SE) for Independence, Exchangeable, AR(1), Unstructured and M-dependent Correlation Structures Estimated Using Unconditional Residuals for GEE and skewed logit-GEE*

Effect	Corr	GEE					SL-GEE				
		Est	Model SE	Rob SE	Wald Z	<i>p</i> -value	Est	Model SE	Rob SE	Wald Z	<i>p</i> -value
Intercept	Ind	0.253	0.228	0.276	0.918	0.359	0.176	0.209	0.239	0.737	0.461
	Exch	0.088	0.263	0.264	0.335	0.738	0.024	0.242	0.235	0.100	0.920
	AR(1)	0.119	0.267	0.276	0.430	0.667	0.043	0.245	0.239	0.180	0.858
	Unstr	0.029	0.272	0.272	0.108	0.914	-0.038	0.249	0.238	-0.160	0.873
	<i>M</i> -dep	0.129	0.261	0.276	0.468	0.640	0.050	0.239	0.239	0.207	0.836
Breastfed	Ind	-0.057	0.108	0.157	-0.361	0.718	-0.062	0.099	0.136	-0.455	0.649
	Exch	-0.022	0.146	0.150	-0.148	0.883	-0.027	0.134	0.134	-0.205	0.838
	AR(1)	-0.052	0.137	0.157	-0.332	0.740	-0.058	0.126	0.136	-0.425	0.671
	Unstr	-0.027	0.149	0.155	-0.173	0.863	-0.031	0.137	0.138	-0.224	0.823
	<i>M</i> -dep	-0.051	0.131	0.157	-0.323	0.747	-0.057	0.121	0.136	-0.416	0.678
BV	Ind	1.086	0.431	0.791	1.373	0.170	1.495	0.348	0.420	3.561	<0.001
	Exch	1.049	0.470	0.802	1.308	0.191	1.475	0.371	0.419	3.524	<0.001
	AR(1)	1.000	0.533	0.910	1.099	0.272	1.494	0.421	0.444	3.368	<0.001
	Unstr	0.901	0.530	0.957	0.941	0.347	1.286	0.440	0.553	2.326	0.020
	<i>M</i> -dep	1.017	0.526	0.926	1.098	0.272	1.513	0.417	0.451	3.353	<0.001
BV:Time	Ind	-0.199	0.091	0.145	-1.376	0.169	-0.275	0.077	0.083	-3.310	<0.001
	Exch	-0.198	0.088	0.144	-1.368	0.171	-0.277	0.072	0.083	-3.340	<0.001
	AR(1)	-0.191	0.107	0.163	-1.176	0.240	-0.280	0.088	0.087	-3.214	<0.001
	Unstr	-0.176	0.099	0.170	-1.036	0.300	-0.246	0.084	0.103	-2.383	0.017
	<i>M</i> -dep	-0.196	0.106	0.166	-1.180	0.238	-0.285	0.088	0.088	-3.227	0.001
HIV	Ind	0.189	0.179	0.309	0.612	0.541	0.222	0.159	0.244	0.910	0.363
	Exch	0.248	0.250	0.299	0.830	0.406	0.273	0.225	0.249	1.096	0.273
	AR(1)	0.216	0.234	0.326	0.662	0.508	0.253	0.208	0.253	1.001	0.317
	Unstr	0.193	0.253	0.323	0.596	0.551	0.232	0.225	0.259	0.894	0.371
	<i>M</i> -dep	0.212	0.224	0.328	0.646	0.518	0.250	0.198	0.253	0.989	0.323
Male	Ind	-0.370	0.117	0.168	-2.202	0.028	-0.382	0.107	0.145	-2.632	0.009
	Exch	-0.358	0.156	0.158	-2.261	0.024	-0.358	0.144	0.143	-2.495	0.013
	AR(1)	-0.359	0.148	0.166	-2.162	0.031	-0.369	0.135	0.144	-2.568	0.010
	Unstr	-0.319	0.159	0.165	-1.937	0.053	-0.329	0.145	0.146	-2.262	0.024
	<i>M</i> -dep	-0.363	0.142	0.167	-2.173	0.030	-0.373	0.130	0.144	-2.589	0.010
Time	Ind	0.178	0.062	0.076	2.346	0.019	0.192	0.057	0.066	2.915	0.004
	Exch	0.146	0.067	0.075	1.941	0.052	0.165	0.061	0.065	2.539	0.011
	AR(1)	0.135	0.071	0.076	1.783	0.075	0.150	0.065	0.065	2.289	0.022
	Unstr	0.110	0.069	0.075	1.457	0.145	0.126	0.063	0.065	1.918	0.055
	<i>M</i> -dep	0.143	0.070	0.076	1.875	0.061	0.156	0.064	0.066	2.370	0.018
Weight	Ind	-0.112	0.057	0.071	-1.581	0.114	-0.125	0.052	0.061	-2.043	0.041
	Exch	-0.068	0.067	0.069	-0.982	0.326	-0.087	0.062	0.060	-1.447	0.148
	AR(1)	-0.062	0.067	0.071	-0.871	0.384	-0.076	0.062	0.061	-1.242	0.214
	Unstr	-0.034	0.068	0.072	-0.481	0.631	-0.050	0.063	0.062	-0.805	0.421
	<i>M</i> -dep	-0.068	0.065	0.071	-0.953	0.341	-0.080	0.060	0.061	-1.311	0.190

after 12 iterations, exchangeability was achieved after 16 iterations, unstructured compliance was achieved after 14 iterations, AR1 was achieved after 89 iterations and after 108 iterations, the model was m -dependent.

The results showed significant differences in the coefficients and their marginal effect, particularly in the interaction terms. When we chose a p -value =0.05 level of significance, parameter estimates from the standard GEE were not significant. In this case, only time and gender were significant. However, when using the skewed logit GEE, gender, time, BV, and the interaction between time and BV were significant.

TABLE 4.11: *Calculated coefficient of bacterial vaginosis with time from $\exp(\beta_1 + \beta_{15} \times \text{time})$, achieved by replacing the respective values from the skewed logit-GEE model with the AR-1 correlation structure*

Time	Coefficient of effects of bacterial vaginosis
Month 1	3.37
Month 2	2.54
Month 3	1.92
Month 4	1.45
Month 5	1.11
Month 6	0.83

4.3.1 Effects of time on BV

This work proceeded and calculated the effects of BV across time on infants given by $\exp(\hat{\beta}_1 + \hat{\beta}_{15} \text{time})$ and reported in Table 4.11. This table shows that the effects of BV on morbidity tend to decrease with time from month 1 to month 5. For example, if we compare month 1 and month 5, this work can conclude that at month 1, the OR of having morbidity incidences are 3.37 times higher for exposed than unexposed babies. At month 5, the OR decreases to 1.11 for exposed babies. At month 6, we observe a reverse causality, whereby the unexposed had higher OR for morbidities.

This can be explained such that sick babies had more hospital visits and therefore, were treated for different illnesses, thus achieving a better health status in the long run. This leaves the BV unexposed group of babies vulnerable to other illnesses during growth, with minimal health intervention as they rarely sought medical attention. This is likely due to the non-threatening nature of the health conditions. With time, these could have led to an increase in illnesses experienced by infants in the unexposed group.

4.3.2 Discussion

In the present study, this work utilized the skewed logit technique under the GEE framework to analyze the risk factors associated with BV. It built on the existing contributions put forth by Nagler [Nagler, 1994] and Liang and Zeger [Liang and Zeger, 1986]. The model adopted in the present study is based on logistic regression, but modified assuming a parameter for skewness, to allow it to accommodate both symmetric and asymmetric responses.

There are several situations in which the relationship between the function of the response and covariates is not strictly symmetric. The asymmetric model is a class of models that borrows strength from both symmetric and asymmetric forms and can be applied in both scenarios, while still maintaining model parsimony. Furthermore, the frequently encountered assumption of symmetry is very restrictive, unrealistic, and can lead to incorrect conclusions regarding the parameter estimates.

The model adopted by this thesis has been shown to be useful in applications when the symmetry properties of a binary outcome are unknown, and it seems to be applicable

in both symmetric and asymmetric cases. Due to the correlated nature of longitudinal data, and needing an easy means of marginal interpretation, the GLM methods seem insufficient, but the use of GEE has been recommended and successfully applied in recent literature.

This work found that gender is a reliable predictor of infant morbidity. Specifically, girls were more likely to be healthy than boys. This finding is supported by previous studies and adds to the large body of knowledge indicating that boys require more attention and health care than girls. With girls having a higher survival probability than boys, our results appear consistent with the reports of Stevenson et al. [[Stevenson et al., 2000](#)]. This finding implies that there is hopes for a decline in mortality among boys if better interventions targeting their health can be implemented.

BV was found out to have a significant relationship with infant morbidities when other covariates are controlled for. Infants whose mothers tested positive for BV were found to have higher morbidity incidences compared to those whose mothers tested negative. The effect of BV on infant health has been reported in several studies, but with different conclusions on morbidities and mortalities [[Lassi et al., 2013](#), [Monebenimp et al., 2011](#)].

The most important finding in this work was the degree of significance observed in the skewed logit model for the interaction between BV and time. This finding would be of interest to doctors, as it indicates the need to plan for proper treatment and monitoring of an infant's health after confirming the maternal BV status, particularly during the first six months. This finding can also inform targeted infant morbidity campaigns, depending on the mother's BV status and the age of the infant.

The negative coefficient of weight and infant morbidities could indicate that an increase in weight gain could reduce morbidity. Babies who eat well tend to gain nutrients from food and have better capabilities of fighting illness in their bodies. Proper weight gain is also an indicator of proper growth. These results were consistent with those reported by [Berger et al., 2007].

Past work by Verma *et al.* indicated an increase in the number of illnesses during infant growth when [Verma and Kumar, 1968]. This is in contrast with what was reported in table 4.9, which shows only an insignificant decline among all the infants(5%), from a high of 61% to a low of 56%. To be more precise, considering the BV-exposed group, there was a huge decline of 25%, from a high of 78% in the first month to 53% in the sixth month.

However, in the non-exposed group, there was a slow increase, whereby the number of morbidities observed increased from 47% in month one to 58% in month six. Moreover, this study was based on the general population of the infants, without factoring in any other factors defining the exposed group or applying any randomization.

Not all covariates included in our study were statistically significant at $p = 0.05$. Nonetheless, their coefficient sign could assist in detecting a trend of association with the response. The covariate set included, e.g. the mode of feeding, whereby breastfeeding had a negative relation with infant morbidities. This finding could reflect behaviors that have been reported in other studies whereby breastfed infants were healthier than their counterparts who were formula-fed [Nduati et al., 2000, Venkatesh et al., 2011].

Finally, the HIV status of the infant exhibited a positive coefficient with infant morbidity.

Infants who tested positive for HIV presented signs of morbidity, consistent with the results obtained by [Venkatesh et al., 2011]. Morbidities associated with HIV were found to increase infant mortality risk according to studies conducted in Kenya [Mbori-Ngacha et al., 2001], Botswana [Shapiro et al., 2007], Cameroon [Monebenimp et al., 2011], and South Africa [Venkatesh et al., 2011].

The novelty of this study is the consideration of a skewed logit model under the GEE framework in health research. This study is one of the few studies that specifically explores the effect of BV on infants across time and considering the HIV status.

Longitudinal binomial data are likely to be observed in numerous health fields where the binary components are correlated. Logit and probit models are widely used for modeling this outcome, which means applying the assumption that data is symmetrical. However, some competing methods for symmetry have been proposed as the logit and probit models do not support skewed binomial responses.

This work has shown that skewed logit-GEE is able to show an association between variables which is not identified by the standard GEE. Accordingly, it fits our imbalanced health dataset better. This thesis has further shown the superiority of the SL-GEE over the standard GEE when asymmetry is assumed. In our approach, the score of morbidities is converted to a binary, with asymmetry in the extreme morbidity cases.

Literature supports an association between BV and morbidity among infants [Mwenda et al., 2021b]. Thus, since skewed logit-GEE has predicted a BV-time interaction, this work concludes that asymmetry is an important factor to consider before choosing the analysis method. It must be appropriately accounted for in analytical models to avoid

biases in final parameter estimates, as has been established in this paper and other works [Coelho et al., 2013, Nagler, 1994, Prentice, 1976, Tay, 2016].

Our research has focused on the commonly neglected 'minor diseases' which have been ignored at the expense of 'major causes' of infant morbidity and mortality [Kellerman et al., 2013, Ladner et al., 2013]. Therefore, this work recommend further research and policies that target infant morbidity on a more holistic level.

4.3.3 Model Diagnostics

This thesis subjected the model to test its robustness using the variance ratio method.

Model-based vs Sandwich-based Variance Ratio

Table 4.12 shows the differences in variances from the model and the Huber sandwich estimate in which this work sought to establish by what factor are they different. This was calculated using

$$V.R = \left(\frac{\text{Robust S.E}}{\text{Model S.E}} \right)^2$$

As expected, and confirmed by our results, the major differences between the model-based and empirical variance occur as a result of the independence correlation structure. The largest differences are in the estimated variance of the BV with the sandwich-based variance ratio. There are differences, but these do not have a notable influence on the variances. They are comparable within the correlation structure.

TABLE 4.12: *Differences in Model-based vs Sandwich-based Variance Ratios for both GEE and SL-GEE*

Effect	Corr	GEE		SL-GEE	
		Est	V.R	Est	V.R
Intercept	Ind	0.253	1.465	0.176	1.317
	Exch	0.088	1.006	0.024	0.944
	AR(1)	0.119	1.062	0.043	0.952
	Unstr	0.029	0.996	-0.038	0.917
	M-dep	0.129	1.115	0.050	0.997
Breastfed	Ind	-0.057	2.099	-0.062	1.893
	Exch	-0.022	1.051	-0.027	0.997
	AR(1)	-0.052	1.309	-0.058	1.169
	Unstr	-0.027	1.085	-0.031	1.022
	M-dep	-0.051	1.433	-0.057	1.274
BV	Ind	1.086	3.364	1.495	1.458
	Exch	1.049	2.915	1.475	1.275
	AR(1)	1.000	2.911	1.494	1.109
	Unstr	0.901	3.262	1.286	1.578
	M-dep	1.017	3.108	1.513	1.172
BV:Time	Ind	-0.199	2.507	-0.275	1.168
	Exch	-0.198	2.690	-0.277	1.331
	AR(1)	-0.191	2.328	-0.280	0.977
	Unstr	-0.176	2.923	-0.246	1.505
	M-dep	-0.196	2.446	-0.285	1.006
HIV	Ind	0.189	2.968	0.222	2.343
	Exch	0.248	1.436	0.273	1.225
	AR(1)	0.216	1.947	0.253	1.487
	Unstr	0.193	1.634	0.232	1.322
	M-dep	0.212	2.145	0.250	1.626
Male	Ind	-0.370	2.057	-0.382	1.845
	Exch	-0.358	1.027	-0.358	0.989
	AR(1)	-0.359	1.268	-0.369	1.128
	Unstr	-0.319	1.069	-0.329	1.001
	M-dep	-0.363	1.395	-0.373	1.234
Time	Ind	0.178	1.498	0.192	1.344
	Exch	0.146	1.261	0.165	1.137
	AR(1)	0.135	1.132	0.150	1.003
	Unstr	0.110	1.207	0.126	1.090
	M-dep	0.143	1.180	0.156	1.049
Weight	Ind	-0.112	1.538	-0.125	1.365
	Exch	-0.068	1.057	-0.087	0.963
	AR(1)	-0.062	1.120	-0.076	0.979
	Unstr	-0.034	1.098	-0.050	0.988
	M-dep	-0.068	1.189	-0.080	1.042

The least variances differences are observed in the AR(1) correlation structure. This supports our choice for using the AR(1) correlation structure for model interpretation. This is because, for a correctly specified correlation, this work expect the model and sandwich errors to be comparable, thus increasing the efficiency in the estimation of the β 's.

4.4 Predictors for Outpatient Care Cost in Kenya

4.4.1 Model Results

A summary of the total cost for outpatient care incurred by the households from the KHHEUS 2018 data are shown in table 4.13. The minimum total cost by household for outpatient care is 0 and 1 KES (0.01 USD) with a maximum spending of 90000 KES (900USD). The mean spend by households was 1141.18 KES (14.41 USD) while for greater than zero was 1811.63 KES (18.11 USD). Very high standard deviation was observed with the data being right skewed. Therefore, to take care of the skewness, there was need to factor it in and the tweedie model under GEE has been found to be a great candidate for this type of modelling.

The output with the QICu as explained in the six models are as shown in the Table 4.1. The model with the lowest QICu was preferred and it this is the model this work will use for explanation of our results.

TABLE 4.13: A summary of the total cost for outpatient care incurred by the households from the KHHEUS 2018 data. Statistics have been recorded for the survey month total cost ≥ 0 KES (all the households) and survey month cost > 0 KES (Those who spend money on outpatient care only). All measurements are recorded in Kenya Shillings(KES).

Statistic	Total cost ≥ 0 by the household	Total Cost > 0 by the household
Minimum	0	1
Maximum	90000	90000
Mean	1141.18	1811.63
Median	170	640
Standard Deviation	3232.73	3921.27
Skewness	8.5	7.05
Characteristic of the skewness	Right skewed	Right skewed

The output with the QICu as explained in the six models are as shown in the Table 4.14.

The model with the lowest QICu was preferred and it was the only one which this work will use for explanation of our results.

TABLE 4.14: Different model outputs with calculated QICu. The model with the lowest QICu is selected as the best fitting model. In our case, Model 1 is selected as the most parsimonious model for predicting outpatient care cost among households in Kenya using the Kenya Household Health Utilization Survey 2018

	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
QICu	976341.2		976874		977759.3		985834		978755		982713.3	
Coefficient	$\hat{\beta}$	<i>p</i>	$\hat{\beta}$	<i>p</i>	$\hat{\beta}$	<i>p</i>	$\hat{\beta}$	<i>p</i>	$\hat{\beta}$	<i>p</i>	$\hat{\beta}$	<i>p</i>
(Intercept)	6.61	0.00	6.59	0.00	6.49	0.00	6.77	0.00	6.37	0.00	6.88	0.00
Age	0.01	0.00	0.01	0.00	0.01	0.00	0.01	0.00	0.01	0.00		
Wealth Index												
Ref (Poorest)												
Poor	0.04	0.64	0.05	0.59	-0.01	0.87	0.04	0.68	0.00	0.98	-0.02	0.85
Middle	0.09	0.32	0.09	0.34	0.00	1.00	0.09	0.29	0.00	0.96	0.02	0.82
Rich	0.41	0.00	0.40	0.00	0.30	0.00	0.41	0.00	0.31	0.00	0.31	0.00
Richest	0.59	0.00	0.58	0.00	0.53	0.00	0.61	0.00	0.53	0.00	0.42	0.00
Marital Status												
Ref (Single)												
Married	-0.04	0.63			0.00	1.00	-0.03	0.76				
Separated	-0.24	0.07			-0.15	0.25	-0.19	0.17				
Divorced	-0.22	0.07			-0.06	0.63	-0.12	0.35				
Education												
Ref (None)												
Primary	-0.25	0.00	-0.24	0.00			-0.27	0.00				
Secondary	-0.41	0.00	-0.38	0.00			-0.44	0.00				
Post secondary	-0.08	0.52	-0.05	0.70			-0.12	0.33				
Sex												
Ref (Male)												
Female					-0.16	0.00	-0.19	0.00				

The best fitting model with the lowest QICu was Model 1 with coefficient and covariates expressed as ;

$$\begin{aligned} \log \mu = & 6.61 + 0.01\text{Age} + 0.04\text{Poor} + 0.09\text{Middle} + 0.41\text{Rich} + 0.59\text{Richest} \\ & - 0.04\text{Married} - 0.24\text{Separated} - 0.22\text{Divorced} \\ & - 0.25\text{Primary} - 0.41\text{Secondary} - 0.08\text{Post-Secondary} \end{aligned}$$

Where

- μ is the expected cost of outpatient care
- Age is a continuous variable
- Wealth index was in 5 different categories (*Poorest, Poor, Middle, Rich, Richest*). *Poorest* was the reference category
- marital status was grouped into 4 categories (*Single, Married, Separated and Divorced*). *Single* was the reference category
- education status was grouped into 4 categories (*None, Primary, Secondary and Post-Secondary*). *None* was the reference category

Since the results are in logarithmic form, this work convert to exponential for interpretation.

Surprising, age factor for the household head was found to be a significant predictor of outpatient care cost. However, there is no much differences in terms of odds. A unit

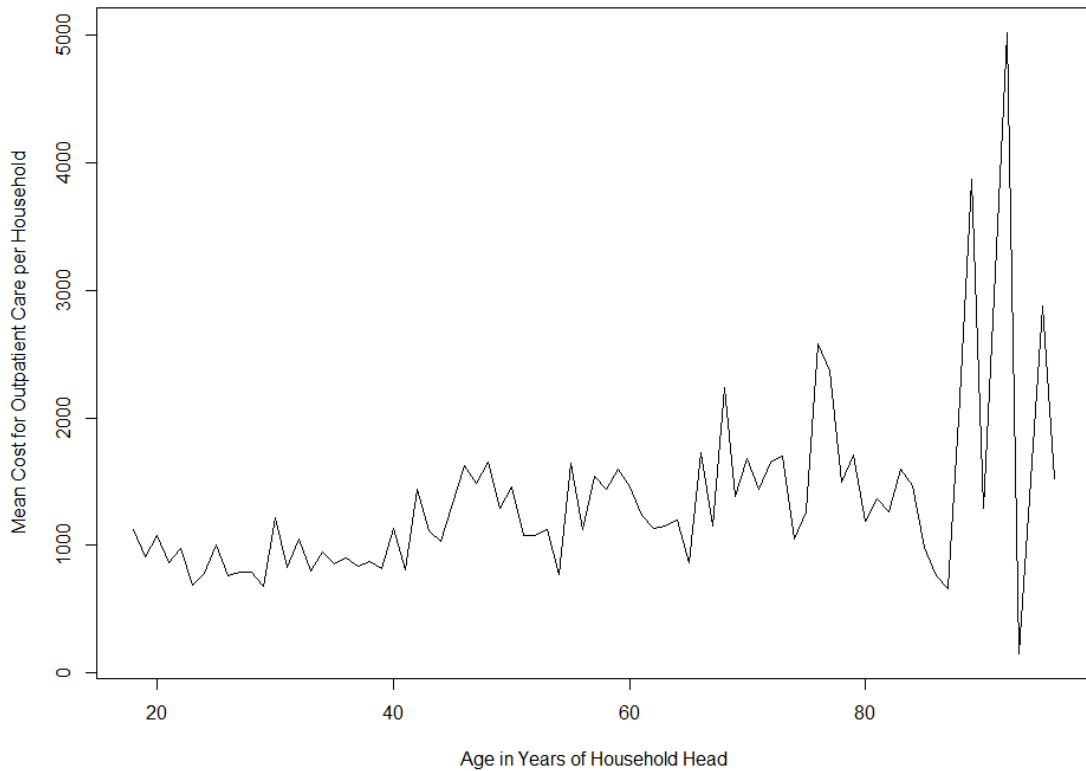


FIGURE 4.2: *Variations of mean cost for out-patient by household head age*

increase with age, results to an increase of healthcare spending by a factor 1.01 (p -value 0.00). The cost of out-patient care was found to change with age in a Sinusoidal manner. Figure 4.2 shows the variation of total cost for outpatient expenses by households with age of the household head during the survey period. Higher cost is experienced where the age of the household head is high.

Outpatient care costs increase across the wealth quantile, with the rich and richest spending more, 1.50 and 1.80 respectively compared to the poorest, results significant at $p=0.05$. The poor and the middle had higher expenses on outpatient 1.04 and 1.09 compared to the poor, but the results are not significant at $p=0.05$.

Outpatient cost varied differently across marital status. Married, separated and Divorced

spend less on outpatient compared to the unmarried 0.96, 0.78 and 0.80 respectively, but not significant at 0.05.

Outpatient cost varied across different levels of education. Primary, secondary and post-secondary spend less compared to those who never attended school at 0.77, 0.66 and 0.92 respectively. However, only primary and secondary were statistically significant while post-secondary was insignificant at $p=0.05$.

4.4.2 Discussion

This study utilizes the Generalized Estimating Equations techniques to analyze predictors of outpatient spending in Kenya. The study develops and uses the Quasi Information Criteria [Hardin, 2013] to assist in identifying factors that best show association of outpatient care with the relevant covariates. This study builds on existing work by [Swan, 2006] and [Dunn and Smyth, 2005].

The models in this paper are based on Quasi likelihood criteria, which means this work doesn't need to specify the full likelihood, but it just need to show how the mean relates to the covariates. There are situations where data are correlated and non-normal meaning the conventional methods could be in appropriate for modelling. The GEEs are class of models that cater for both correlated and skewed responses.

Furthermore, the most regularly encountered assumptions of specifications of full likelihood are solved. Consequently, the ability to compute the QICu for model comparison gives us more flexibility in statistical analysis.

In this study, wealth index, as a measure of socio-economic status was found to be a good predictor of outpatient spending. This finding is consistent with literature and has been reported in other studies [Girma et al., 2011, Kevany et al., 2012]. Yet other studies have not shown any association of wealth index and spending since the study population was homogeneous poor [Ngugi et al., 2017]. This implying that spending in relation to wealth index varies by sample selection.

The rich are also likely to seek care in private hospitals which are more expensive, thus the higher costs reported. The Government of Kenya is making efforts to reduce poverty among its citizen in order to raise their socio-economic status to ensure that the households have the extra income to spend on their care. In addition, the government should strive to lower fees for outpatient care to encourage citizens to visit when sick before their conditions deteriorate. Some developing countries have conducted a cost benefit analysis of their healthcare [Shon et al., 2018].

A study by [Muriithi, 2013] found out that user fees was among the key determinants of whether a patient is treated or not. The implication thereof is that, we could be having people who don't seek care during sickness or injury due to the un-affordability of fees charged at the facilities. The wealthy could spend more, since they are spoilt for choice and can afford any doctor of their choice, especially the private clinics. Very minor cases can be assigned to highly trained doctors thus not making full utility of this doctor who could be handling more complicated cases.

The unmarried spend more on outpatient healthcare compared to married, separated or divorced. possibly, they could be spending more on inpatient. For example, this group is mostly comprised of parents, and therefore could be spending more on other illnesses and

services that comes with lifestyle and old age. for example, most chronic diseases affect mainly the aged and services such as child birth are mostly inpatient.

So this finding could suggest that the unmarried spend more on outpatient while the other group could be spending more on inpatient. The finding could caution the youth on what to expect as they move to their next stages of life. The unmarried are mostly younger, and sole decision makers. They are also mostly young and could be trying to be financially stable.

It is interesting to observe that the less educated spend more on outpatient care than the ones who have any education. The learned could be self-medicating for less serious illnesses hence spend less, yet the less educated could only be believing that every illness require hospital visit.

In Kenya, most outpatients are considered out of pocket spending. Most educated could have better financial status and afford different insurance which pay for them. Sometimes the values paid since they are swapped, may not have been captured in the data collection. Most educated have better lifestyle, due to better finances thus can have better health. Better health means you need less outpatient care. In contrast, less educated could mostly have poor lifestyle, poor feeding, poor housing mostly live in slums and have a higher probability of exposure to different illnesses, that would finally require care. This means they have to seek medical care, thus spend more.

Households with elderly persons spend more thus explaining an increase in cost spend with age. As you age, your immunity becomes compromised, more likely to have diseases thus would be required to pay more on outpatient care.

This work went ahead and calculated some probabilities based on [Dunn and Smyth, 2005] to demonstrate the usefulness of the tweedie in modeling cost for outpatient care. When $1 < p < 2$, then the tweedie parametres (μ, p, ϕ) can be parameterized to poisson and gamma parameters $(\lambda, \gamma, \alpha)$ which can be used to provide some estimates that this work can compare to other outputs. This are given in equation

$$\lambda = \mu^{(2-p)} / \phi(2 - p)$$

$$\gamma = \phi(p - 1)\mu^{(p-1)}$$

$$\alpha = (p - 2)/(1 - p)$$

Where λ is the average spend per month, γ is the shape of the cost distribution when a households pays for inpatient care and $\alpha \gamma$ is the mean spend per month.

Considering our best fitting model, the parameter index p is 1.68, $\mu=6.76$, ϕ is 30.85. When it reparametrize to gamma and poisson gives the predicted mean cost spend per month calculated as

$$\lambda = \frac{6.76^{(2-1.68)}}{30.85(2 - 1.68)} = 0.18$$

and

$$\gamma = 30.85(2 - 1.68)6.76^{(1.68-1)} = 36.12$$

finally

$$\alpha = \frac{1.68 - 2}{1 - 1.68} = 0.47$$

The mean amount spend per month on out patient care is $\alpha\gamma = 0.47 * 36.12 = 16.97$ KES (0.17 USD) per month, which translates to 1403 KES (14.03 USD) per year.

Following [Dunn and Smyth, 2005] the probability of incurring zero cost on outpatient by the household (in other words, the probability of not seeking outpatient care) is given by

$$\Pr(Y = 0) = \exp(-\lambda) = \exp\left[-\frac{\mu^{2-p}}{\phi(2-p)}\right] \quad (4.2)$$

such that, probability of zero outpatient is given by $\exp(-0.18) = 0.83$ meaning that 83% of household will not spend any cost on outpatient care in any given month. Therefore only 17% will spend on out patient cost. consistent with other results.(cite khheus)

TABLE 4.15: *The residual deviance and degrees of freedom for a Tweedie glm with differing link functions using Model 1 covariates*

Link function	Deviance	DF
Logarithm	404663.6	11118
Canonical	404872.7	11118

Conclusions

In terms of model selection using the QICu approach, this work selected model 1 as the best model for predicting outpatient care with QICu of 976341.2. However, in terms of the most parsimonious model with the least number of covariates for predicting outpatient expense is model 2 with a QICu of 976874.

This can be explained as follows; adding marital status to the model 2, lowers the QICu significantly, but it is not significant at $\alpha = 0.05$. The differences in the QICu is basically due to penalty imposed during calculation equivalent to increase in the number of covariates.

The data collected were on recall of the 4 visits during the year. In terms of recall, spending on the 4th visit could be more accurate, however our study focused on the first visit. A further research on each individual visit would be necessary.

More spending on subsequent visits was expected since a revisit would mean the previous one was not effective and thus require more medical tests. Also, more research on subsequent visit is required.

There is a thin line between inpatient and outpatient care. A difference in financial status could bring out the difference. For example, a headache could be a symptom of a serious condition like migraines. Having financial strength would facilitate more test like scans which can bring out the issue clearly.

This could lead to an admission for better treatment and probably inpatient admission. However, the poor may opt for pain killers which is what they can afford. In this case, the rich could have its problem solved, while the poor person's health condition would continue deteriorating. This means, a serious inpatient case can be converted to outpatient care due to financial constraints.

In addition, outpatient in Kenyan could also include those who bought medicine at a chemist. So, some of the outpatient could be self-diagnosed. Those who are educated are better placed to do self-diagnosis before moving to a hospital. This could further

our support for less spending during visit 1 by those who are educated than the less-educated. It would be interesting to see how education associates with outpatient care in the subsequent visits, which are more serious.

It is common knowledge that not all the rich are better educated, but most better educated have better socio economic status. However, this work observed the rich spend more, while the most educated spend less, thus the characteristic not coming out clearly. However, the differences could be (1) mode of payment for example, it is easier to recall a cash payment for a service than when someone swipes an insurance card.

This value may be missed during data collection, and (2) computation of wealth index which involve assets which may be outdated, for example a teacher who lives in an urban area far away from his main family could not be having assets like a fridge, TV and could be classified more poorer than a poor household which has a car and telephone which is not functioning.

In this work, the following 17 assets were used to determine how wealthy the households were. Radio, TV with Free to Air Set-top-box/Digital TV, TV with Pay TV Decoder, Internet protocol TV (IP TV), Analogue TV (With no connection/signal), Internet through mobile phone/Modem, Fixed Internet at home e.g. Fiber, Satellite dish, LAN, Wi-Fi, Computer/Laptop/Tablet, Bicycle, Motor Cycle, Car, Truck/Lorry/tractor/ bus/ three wheeler truck, Refrigerator, Motor boat, Animal Drawn cart, Canoes and Tuktuk.

The challenge with this approach is that a household which has a Digital Tv that combines functionality of Radio and TV worth like 5000USD, is ranked poorer than a household with a detached Radio, digital TV and analogue TV not in use but available, worth

20USD, 500USD and 200USD. The sum which is much lower than the single TV held by the different household. Computation of such should be revised. Additionally, it would be interesting to see if this characteristic is also observed on inpatient care cost.

This study recommends UHC as it will be an equalizer for all the Kenyans who need medical care as well as other developing countries.

Finally, a reproducible R code is provided in Appendix 9.

Further analysis on the Age

Also, the single best predictor of cost was age with a QICu of but although wealth index had a lower QICu. The same principle of penalty applies and the difference in QICu is ignored. So, this work investigated further the age variable that can inform policy.

Include spatial maps to make the thesis cute (plot both median age and median spend and mean and) The variable age was statistically very significant, but the coefficient 1.01 showed no major difference. However, this variable was wiggling and behaved in a sinusoidal manner.

It also had the lowest Qicu. In order to draw a valid conclusion on its implication, this work subjected it to bivariate spatial mapping to understand how it varies regionally with cost. This work plotted age versus mean spending in Kenyan counties.

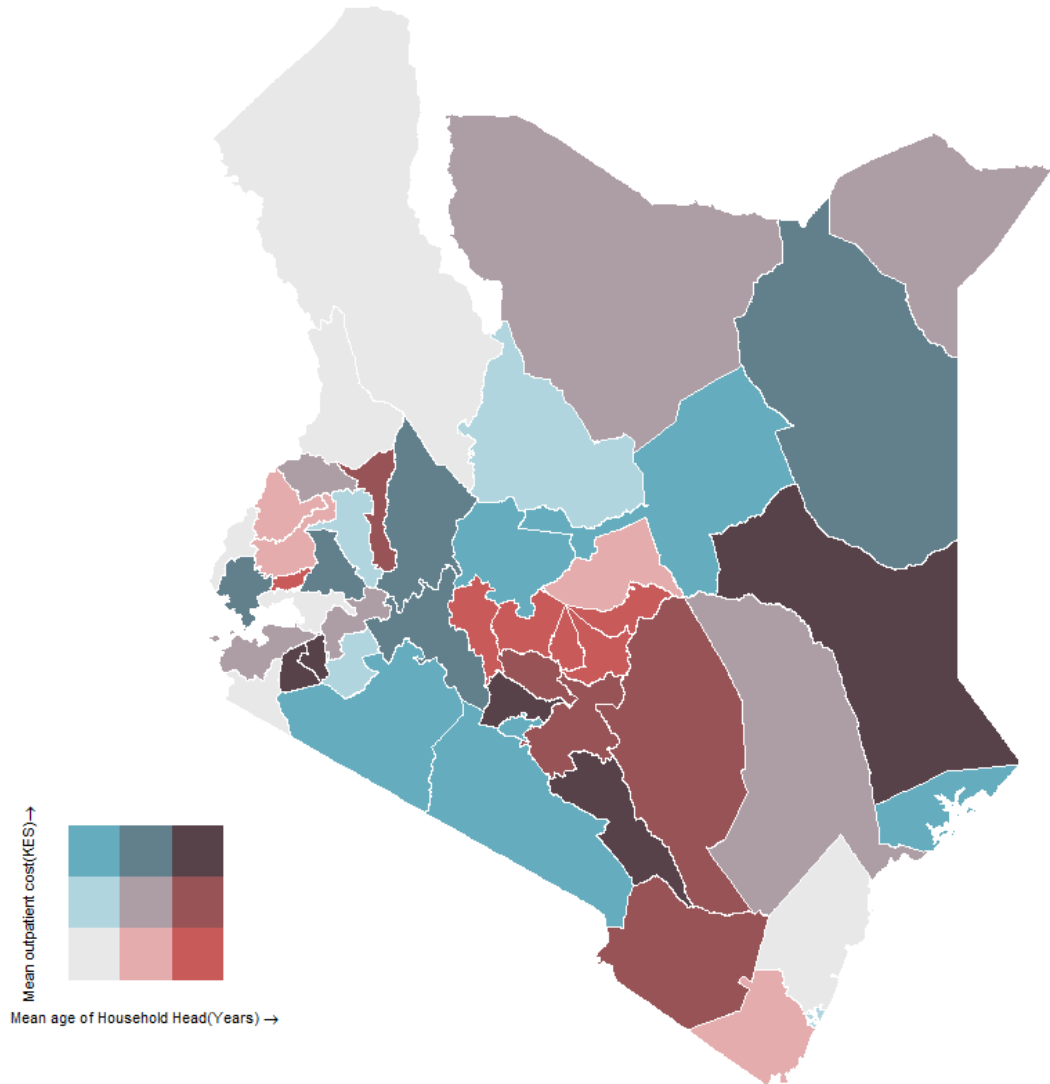


FIGURE 4.3: *Bivariate plot of mean cost for mean out-patient costs by household head age for regions in Kenya*

Interpreting the bivariate map of Mean outpatient cost and age of the household head

In our map, this work identified that in our map 4.4, we identified that counties (Turkana, West-Pokot, Busia, and Migori) in North western part of the country, (Kilifi) in the South eastern part of the country fell in Lowest household head age/Lowest outpatient cost suggesting that this counties may be important targets for increase in resources since

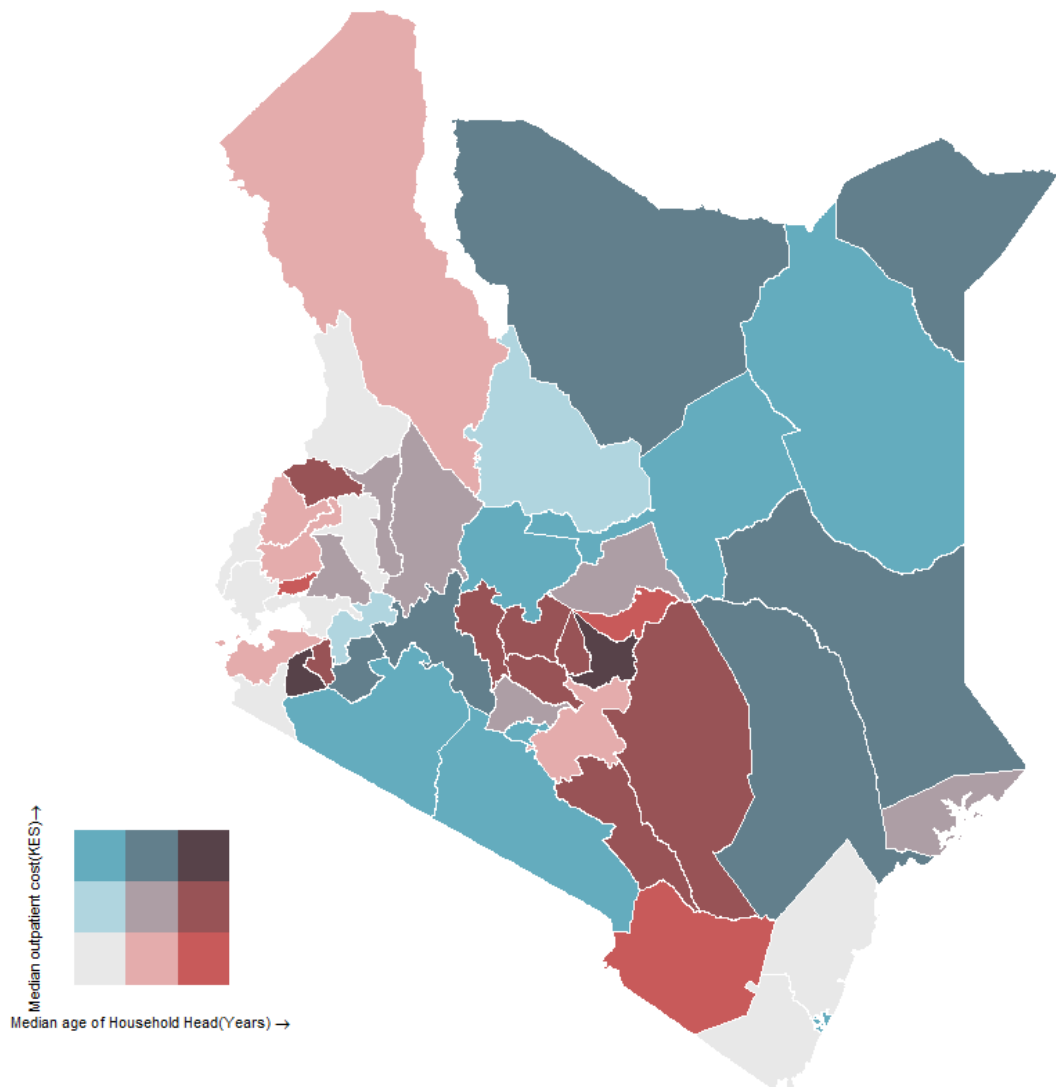


FIGURE 4.4: *Bivariate plot of median cost for out-patient costs by household head age for regions in Kenya*

they are also among the poorest.

This counties are headed by young heads of households thus policies targeting having more responsible members should be developed. Additionally, counties with high household age and high cost of outpatient (Kisii and Nyamira) in western. Kiambu in Central and Makueni in Eastern may be important for targeting reduction in cost for outpatient care and inequalities in wealth.

Counties with High age/low mean cost (Nyandarua, Nyeri, Muranga, Kirinyaga, and Embu) in central could suggest could have more population in old age

High cost/low age (Narok and Kajiado) in southern, (Laikipia and Isiolo) in Eastern and Lamu in South eastern would require policies targeting the lowering cost as they are the largest counties in terms of land size and cultures.

To reverse the inequalities, this thesis suggest a shift in distribution of resources across counties.

4.5 Results from Analysis of Distance Traveled for Inpatient Care in Kenya

4.5.1 Model selection

This work presented 10 competing models for distance that demonstrates the best-fitting model with the lowest QICu, as shown in Table 4.16. This work used the backward selection approach [Zhang, 2016] as a proxy to identify the best predictors for distance under a GLM.

However, our model output and interpretation are only based on distance adjusted for the respective covariates in the GEE framework. This work added the covariates into the model and computed their QICu and R^2 . It then removed the covariates one by one and checked whether the changes improved the model fit. The model with the best-fitting covariates from Table 4.16 is model 7 written as

$$\log(\mu) = 0.093 + 0.222\text{paidMed} + 1.226\text{paidHigh} - 0.523\text{employed} + 0.092\text{hMed} + 0.471\text{hLarge}$$

where μ is the expected distance traveled to access inpatient care. In this model, the amount paid for healthcare (medium and high) takes a value of 0 or 1 depending on which category is being assessed.

The low group is the reference category. Employment takes a value of 1 if the respondent is employed. Finally, household size (medium and large) takes a value of 0 or 1 depending on what is being assessed. The reference category, small household size, does not appear in the equation. This model resulted in $\phi=6.12$, $\alpha=0.045$, $R^2=9.96\%$, and $\text{QICu}=13158.23$ with $p=1.64, 95\%$ CI(1.59,1.68).

TABLE 4.16: *Models selection using QICu and R²*

Model number	Covariates	QICu	R ²	Variance power P(95%CI)	No. of co- variates
10	Ability to pay, employment status, household size, wealth index, education level, age	13304.7	10.39	1.63(1.58,1.67)	6
9	Ability to pay, employment status, household size, wealth index, education level	13306.16	10.41	1.63(1.59,1.67)	5
8	Ability to pay, employment status, household size, wealth index	13317.38	9.7	1.62(1.59,1.67)	4
7	Ability to pay, employment status, household size	13158.23	9.96	1.64(1.54,1.68)	3
6	Ability to pay, employment status	13280.7	9.5	1.64(1.59,1.68)	2
5	Have insurance, place of residence	12733.1	0.19	1.67(1.63,1.71)	2
4	Ability to pay	13066.33	8.38	1.64(1.60,1.68)	1
3	Place of residence	12773.2	0.17	1.67(1.63,1.71)	1
2	Household size	12755.65	0.54	1.67(1.63,1.71)	1
1	Employment	12698.31	1.4	1.67(1.63,1.67)	1

Model 7 was selected as the best model, even though models 9 and 10 had higher R² values. This work selected the model with the best balance between the QICu and the R², in which model 7 fits as the best parsimonious model, with the least covariates with acceptable QICu and R² values.

To interpret the coefficients, which are captured in logarithmic form, the exponential was taken. From the given model, all factors remained constant, and the population average

distance to a government inpatient center in Kenya is approximately $\exp(3.093)$, which is 22.04 km.

Compared to those who paid the lowest amounts for healthcare (1–3,000 KES), citizens in the middle pay category (3,001–10,000 KES) traveled 1.24 times the distance to a healthcare facility, whereas those who paid the most traveled 3.40 times the distance.

The employed traveled half the distance to a healthcare facility for inpatient care than the unemployed (0.59 times). Compared to small household sizes (1–3 members), medium households (4–6 members) traveled 1.096 times the distance to a healthcare facility, and the largest household sizes (7+) traveled 1.60 times the distance than small households.

Finally Table 4.17 shows under the current setting, considering a Logarithmic link function is more acceptable since it has lower deviance compared to the canonical link function, which is the default in the software.

4.5.2 Discussion

This work has demonstrated the use of a new technique for clustered and correlated non-normal responses that depict a discrete mass at zero under generalized estimating equations. This study presents the best set of covariates for predicting distance traveled by Kenyans to access inpatient care from 47 counties. Data from each county are representative, and the pooled data contributed substantial information about distance for inpatient care. A set of potential covariates was investigated to better understand their effects.

TABLE 4.17: *The residual deviance and degrees of freedom for a Tweedie glm with differing link functions using Model 7 covariates*

Link function	Deviance	DF
Logarithm	5251.198	455
Canonical	5280.799	455

The model without covariates showed that, on average, a Kenyan seeking inpatient care traveled a distance of 22.04 km. However, the travel cost can differ substantially, in that the road terrain, which is preferred for accessing hospitals, varies widely in Kenya. Some roads are all-weather, whereas others are seasonal, meaning during rainy times, accessibility is greatly hampered.

Healthcare system performance can be assessed according to the healthcare service distribution, access, and utilization [Thaddeus and Maine, 1994]. Access is mostly determined by cost and distance. Thus, irrespective of the availability of a service in a hospital, if it is not utilized by the target group, its full utility cannot be actualized. The aim of the United Nations Sustainable Development Goal 3 is to ensure healthy lives and promote well-being for all at all ages, and this work shows the importance of distance in measuring this goal.

Most inpatient care is usually critical and requires specialized attention by a medical expert; therefore, distance to access could determine survival. Although some studies have not linked accessibility to use [Nesbitt et al., 2016], there has been evidence that ease of access could potentially save lives, as some life-threatening conditions are worsened by long distances to see a physician. For example, when a patient suffers a heart attack, the time to get to the hospital can determine their survival.

As shown in the inpatient data, some patients were admitted to healthcare centers operated by the government. Although not able to handle extreme cases, they are well-equipped to handle many cases, such as childbirth. However, in the event of complications from childbirth, such as the need for cesarean services, patients may need to be referred to a larger facility. Therefore, major conditions still need referrals to large hospitals, and thus access for inpatient care for these services at lower-level hospitals remains a challenge.

[Noor et al., 2006] investigated access to government healthcare centers and reported a distance of less than 10 km. Although they focused on only four districts in Kenya, these districts could be used as a proxy for the national distance estimate.

However, their focus was on small healthcare centers mostly used for outpatient care and thus could be misleading when predicting inpatient access. Also, it is important to note that distance in Kenya is difficult to predict because of the differences in terrain and road types (e.g., tarmacked, marram grass, and earth); thus, a low value R^2 of 9.9% is reasonable.

Distance for inpatient care is important to the Kenyan government as it tries to achieve universal healthcare coverage. The main goal of universal healthcare is to ensure that every citizen has access to quality healthcare services; however, this can only be achieved if the distance to access is reasonable and achievable. What this means is that for the government to achieve the general objective, it needs to improve access for inpatient services.

It is also evident that although distance to inpatient services is generalized, Kenya has

a unique geographical terrain that can affect access. For example, it may be easier to access a facility that is further away in an urban area than close by in a rural area. This is because most urban cities have good road networks, making access easy. If the government wants to increase access, more needs to be done to improve the road network infrastructure in rural areas.

Our results show that high costs are associated with longer distances traveled to access inpatient care. This work can interpret this in two ways. First, the cost incurred could signify an expensive procedure or care. Second, those with higher incomes could choose a facility that is farther away and more expensive, even though the required care is not complicated, as they could be more confident regarding the care in these hospitals.

Those who had paid the most (10,001+ KES) tended to travel a greater distance (up to 3.40 times the distance) for inpatient care compared with those who paid the least (1–3,000 KES), and those in the middle amount paid category (3,001–10,000 KES) tended to travel 1.24 times the distance. Higher wealth gives a person the freedom to choose any facility they are comfortable with for inpatient care, and high hospital fees are associated with complex medical needs and procedures.

For example, a cesarean section costs more than a normal delivery, although both require inpatient care. However, a large hospital is likely more suited to handle a cesarean on a woman with preeclampsia than small hospitals. This is because sophisticated medical equipment is required for the procedure and is mainly found in large referral hospitals. Therefore, a patient with high financial means will travel longer distances for the procedure and pay higher medical costs. Those with low incomes will check in at the closest and most affordable facility.

Our results indicate that Kenyans traveled longer distances for complex medical procedures or for better services, which may not necessarily be found in the closest inpatient healthcare facility. Some people may have traveled a long distance to obtain privacy. For example, a person could be more comfortable being admitted for inpatient care for an STD in a hospital further away from home. However, our results may not be comprehensively conclusive, and further investigation on why those who travel farther paid more for services should be established.

In Kenya, after adopting a new constitution in 2010, healthcare services devolved to obtain proper and closer management at low levels. However, due to limited budgets and other indirect effects, such as poor roads and lack of electricity, there has been slow growth in terms of hospital upgrades for inpatient care. For example, setting up an inpatient care facility deep in a rural area without a good road network or proper supply of electricity would be meaningless in terms of serving the people. Thus, facilities are mainly established in areas where such services can be accessed easily. This means that people in remote areas still must travel long distances for inpatient service.

Inequalities in employment opportunities also determine the distance traveled to access inpatient care. It is evident that the employed travel half the distance as the unemployed to obtain care, as supported by [\[Allin et al., 2009\]](#).

This means that the unemployed (with lower incomes) are forced to use facilities within reach and may be prevented by the lack of financial resources to access better facilities for specialized treatment. Additionally, the employed have the advantage of being able to afford to live in an area where large inpatient facilities are found.

For example, large referral and specialized hospitals are typically found in capital cities and large towns so they are accessible and serve many people. These facilities also need to be connected to an uninterrupted supply of water and electricity. Most employed people choose to live in places where such services can be found.

Family size was the last covariate that determined distance traveled, with medium and large households traveling longer distances than small households. This difference could be because large households are mostly found in rural areas and slums rather than urban areas. It is easier to raise a large family in rural areas because food and accommodations are affordable, as many people live on their ancestral land where they also farm most of their food.

This shows that there is a need to improve inpatient facilities in rural areas and slums. Without a strong policy focus to support equal access to inpatient services in Kenya, prioritizing the rural areas and slums, opening up job opportunities, and encouraging smaller families, the dream of achieving universal healthcare coverage will remain unfulfilled.

This work is the first to estimate the distance for inpatient care in Kenya, analyzing all responses from 47 counties. This provides the best estimate and evidence on which policies to formulate. This area has been understudied by researchers focusing on both inpatient and outpatient care, and an analysis of the scarce existing literature could lead to wrong conclusions and poor policy formulation.

For example, [Noor et al., 2006] reported a distance to access of less than 10 km, which indicates that every person has access to healthcare. However, as stated earlier, most of these facilities are for outpatient care. Thus, if policies are based on this conclusion, there

may not be improvements to much-needed inpatient care that requires sophisticated and complex procedures, as well as doctors.

A drawback of previous studies is the analysis of distance using summary statistics, whereby researchers do not dig deep into the data and only report averages. Using Table 2 as a guide, for example, this work could have reported a median of 10 km, which may be misleading, as it did not factor in skewness and the correlations that exist in the data, which could provide more insight. This shows that our advanced statistical analysis provides a more meaningful interpretation of the data by factoring in both skewness and correlations.

We calculated probabilities based on [Dunn and Smyth, 2005] to demonstrate the applicability of the tweedie distributions in modeling distance traveled to access inpatient care. When $1 < p < 2$, then the tweedie parameters (μ, p, ϕ) can be parameterized to poisson and gamma parameters $(\lambda, \gamma, \alpha)$ which this work uses for estimation. This are given in equation

$$\lambda = \mu^{(2-p)} / \phi(2-p)$$

$$\gamma = \phi(p-1)\mu^{(p-1)}$$

$$\alpha = (p-2)/(1-p)$$

Where λ is the average distance traveled, γ is the shape and α γ is the mean

Considering our best fitting model, the parameter index p is 1.64, $\mu=22.04$, ϕ is 6.12.

When this work reparametrize to gamma and poisson gives the predicted mean cost

spend per month calculated as

$$\lambda = \frac{22.04^{(2-1.64)}}{6.12(2-1.64)} = 1.38$$

and

$$\gamma = 6.12(1.64 - 1)22.04^{(1.64-1)} = 28$$

finally

$$\alpha = \frac{1.64 - 2}{1 - 1.64} = 0.56$$

The mean distance to access inpatient care is $\alpha\gamma = 0.56 * 28 = 15.75$

Following [Dunn and Smyth, 2005] the probability of incurring zero cost on outpatient by the household (in other words, the probability of not seeking outpatient care) is given by

$$\Pr(Y = 0) = \exp(-\lambda) = \exp\left[-\frac{\mu^{2-p}}{\phi(2-p)}\right] \quad (4.3)$$

such that, probability of households covering zero distance is given by $\exp(-1.38) = 0.25$ meaning that 25% of household that require inpatient care will cover less than 2KM; such that three quarter of the population cover large distance to access inpatient services. (only 25% of households had inpatient care within their reach).

This thesis have demonstrated a new approach for handling correlated non-normal data and created an R function GitHub repository

<https://github.com/samwenda/Tweedie-with-Exchngable-Correlation> Our approach

has demonstrated how distance decay can affect access to much-needed healthcare services. Our approach can be used in other datasets that have a discrete mass at zero and correlation within clusters. Our study has also provided a new way of calculating the denominator, following [Hardin, 2013] as shown in Appendix 2.

Further advancement of this study could be to focus on individual analyses of the 47 counties. This is out of the scope of this work, which only focuses on the population average. Although some studies have found no association with use [Nesbitt et al., 2016], better access regardless of the quality is still influenced by distance.

This study has some limitations. First, the data were missing a significant amount of information, which made it difficult to input. Therefore, this work only used complete data. New complex statistical methodologies for predicting non-normal data need to be developed.

Finally, for policy implications, policies targeting having more government healthcare facilities in rural areas and slums should be developed because large populations exist in those areas.

In addition, policies advocating smaller families should be encouraged to ensure that people can afford better healthcare. Job availability will also increase flexibility in the choice of a facility. Sophisticated services should be brought closer to low-income families to ensure they do not travel long distances for much-needed services.

4.5.3 Model Diagnostics

4 competing models were subjected to further analysis to find the most appropriate model.

The models are 7,8,9 and 10.

4.5.4 A non-parametric test of the randomness of residuals

[Hardin, 2013] has documented clearly the test to be conducted as advised by [Chang, 2000] who directed using the non parametric test such as the Wald-Wolfowitz randomness test. This analysis determines if the model assumptions are violated by the data.

The Wald-Wolfowitz randomness test, is Z-statistic given by equation 4.6, which was performed on each of the 4 models to test the signs of raw residuals were distributed in a random sequence.

Following [Hardin, 2013], the expectation given as

$$E(T) = \frac{2n_p n_n}{n_p + n_n} + 1 \quad (4.4)$$

and Variance given as

$$Var(T) = \frac{2n_p n_n (2n_p n_n - n_p n_n)}{(n_p + n_n)^2 (n_p + n_n - 1)} \quad (4.5)$$

The test statistic for the hypothesis that the signs of the residuals are randomly distributed is,

$$W_z = \frac{T - E(T)}{\sqrt{Var(T)}} \quad (4.6)$$

which has an approximately standard normal distribution, and thus the corresponding p-value can be determined using z-tables one sided.

TABLE 4.18: *Models Diagnostics using Wald–Wolfowitz run test*

Model	Negative Residuals (n_r)	Positive Residuals (p_r)	Observed runs (T)	Expected Value($E(T)$)	Variance($Var(T)$)	Test Statistic (W_z)	p-value
10	330	121	179	178.07	69.28	0.111	0.456
9	336	115	177	172.35	64.87	0.577	0.288
8	336	115	169	172.35	64.87	-0.416	0.338
7	332	119	165	176.20	67.82	-1.36	0.086

Extreme values of Wald–Wolfowitz run test(W_z) indicate that the model does not adequately reflect the underlying structure of the data

The test reveals that using $\alpha=0.05$ there is not enough evidence to reject the hypothesis that the residuals from the model are random. In general, the result of the runs test does not significantly change due to the hypothesized structure when the model is correct in terms of including necessary covariates in their proper form. Our result suggest that since the 4 models meet the minimum thresh hold for randomness, then they can be used as predictors for distance of inpatient care. It is now at the discretion of a researcher to select the models based on literature for analysis and prediction.

4.5.5 Raw residuals analysis using Plots and graphical assessments

Since the 4 competing models passed the threshold of Wald–Wolfowitz run test, the other approach for checking model adequacy is to investigate plots of the raw and Quantile residuals.

Raw residuals versus the observation numbers

The plot shows by Figure 4.5 shows that the magnitude of the positive residuals is larger than that of the negative residuals. This indicates that all the four models are not sufficient enough to model extreme cases of distance as accurate as possible. In addition, distance travelled for inpatient care had to be greater than zero thus the negative residuals are expected to have a lower magnitude. With this results, this work prefers to subject our models to further tests.

QQ normal Plots

Another criteria suggested by researchers for testing randomness is the QQ plots. The plots as shown by Figure 4.6 shows that all the model are appropriate for modeling distance for inpatient care, as the residuals lie close to the Normality line. This means with such results, this work still need further testing to determine the best model fit.

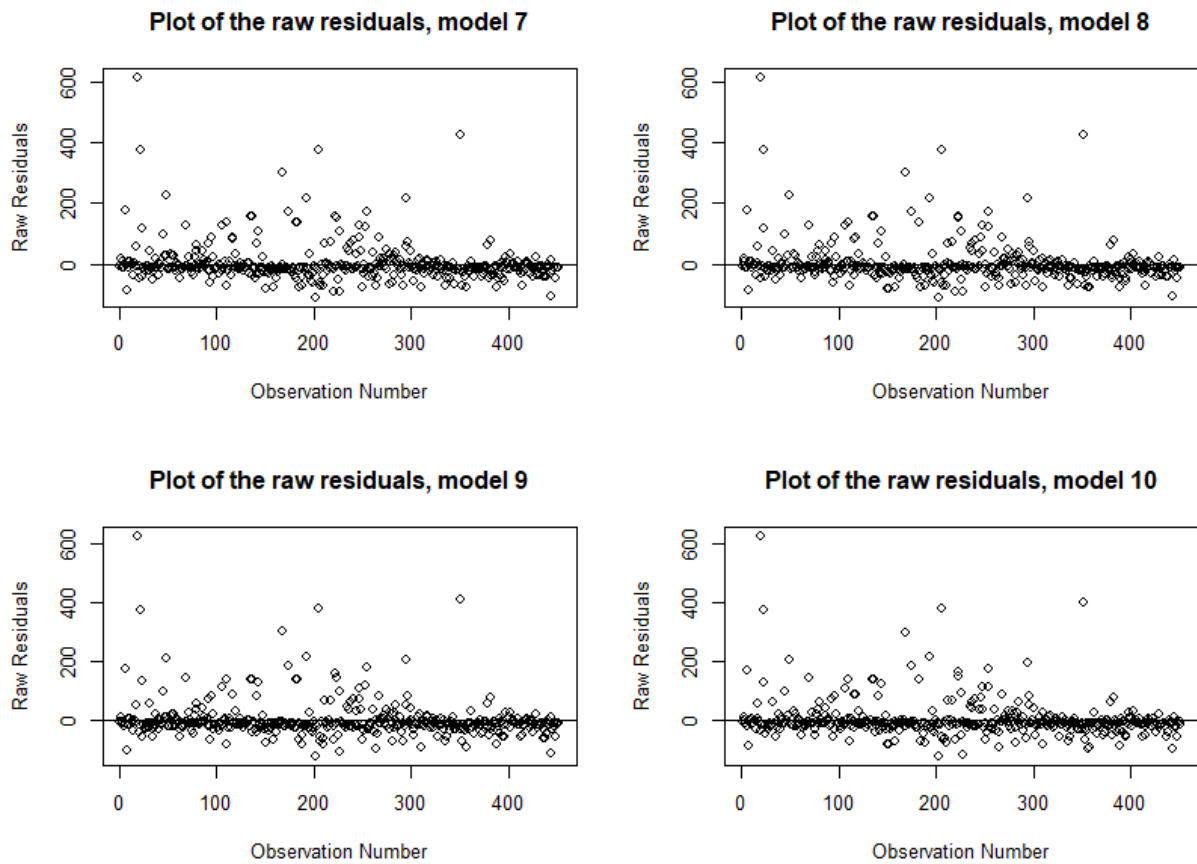


FIGURE 4.5: *Plot of raw residuals versus the observation numbers*

Raw residuals versus the Linear Predictor

It has been established that there were no differences in terms of graphical representation on for both tests, that is the QQplot and the raw residual versus observation number. This work therefore subject the models to further test to try seek if it can find the model that fits the data well. This work considered the raw vesus linear predictor plots as shown by Figure 4.7

This plot gives a great direction in terms of model selection. It is evident that model 8,9 and 10 have similar fit and do not show uniform spread of the points. This means the models have a poor fit in predicting distance as far as residual plots are concerned.

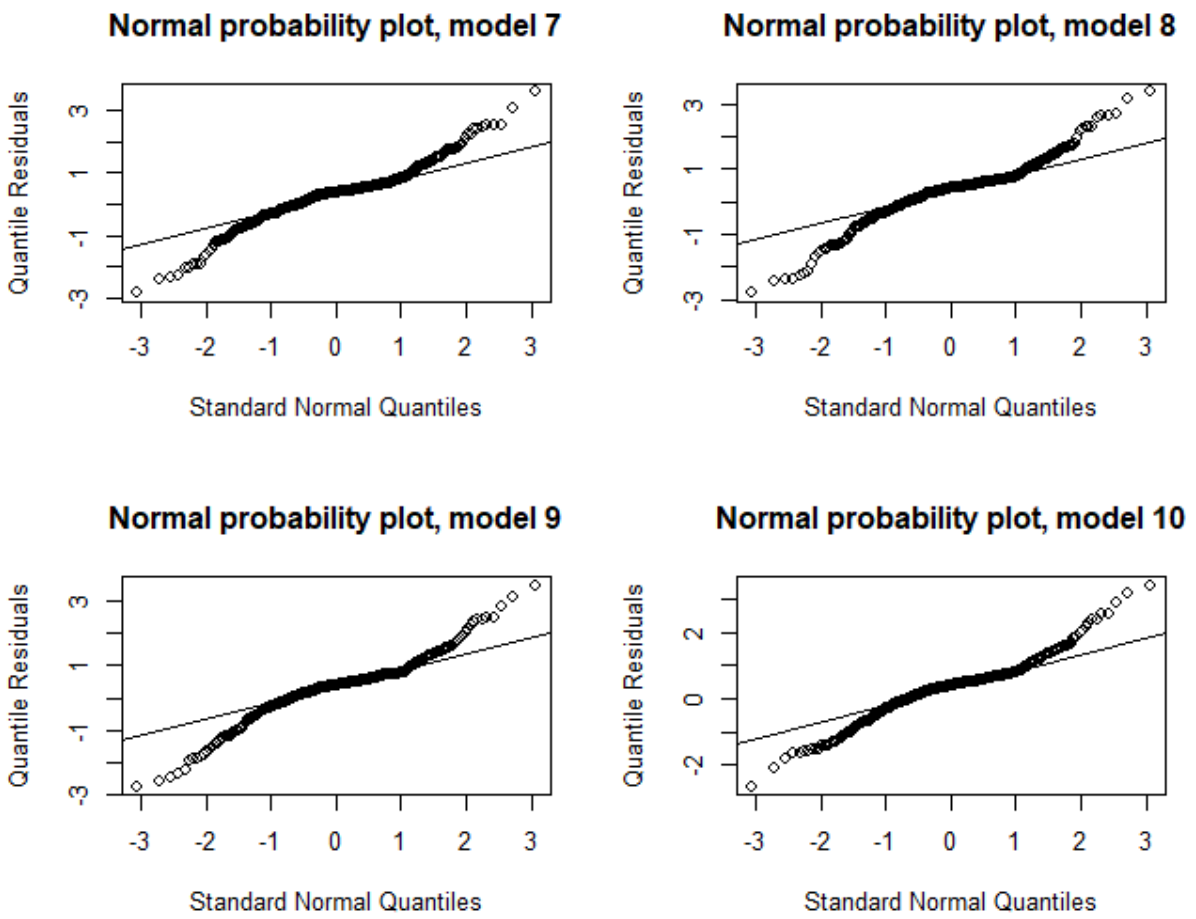


FIGURE 4.6: *QQ normal Plots*

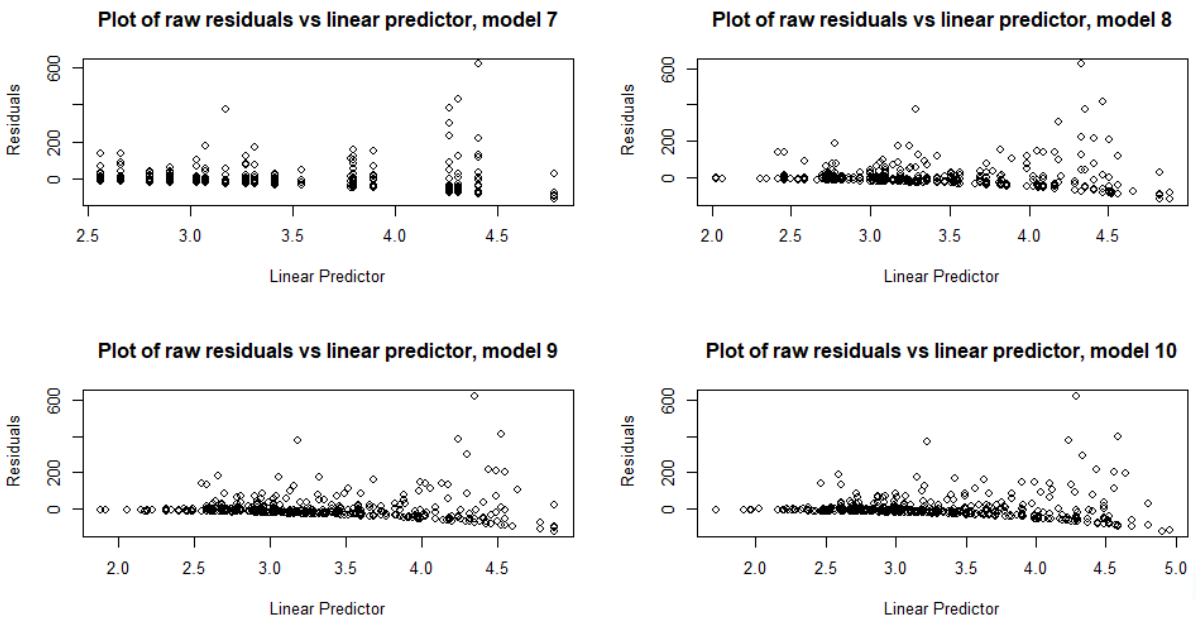


FIGURE 4.7: *Plot of raw residuals versus the Linear Predictor*

However, model 7 has better spread of points showing a better fit, and is our most preferred model.

Chapter 5

Conclusions, Recommendations and Further Research

Conclusions

This thesis has investigated various aspects of non-normality and correlation under the GEE framework. Past application under non normality assumptions have seen utilization in; (1) insurance field on modeling claims [[Peña-Sanchez, 2019](#), [Smyth and Jørgensen, 2002](#)]; (2) meteorological field on analyzing rainfall [[Hasan and Dunn, 2012](#), [Swan, 2006](#)]; (3) Health field on examining medical costs [[Kurz, 2017](#)] and (4) Dose response data [[McDaniel et al., 2013](#), [Prentice, 1976](#)].

We evaluated two types of datasets, (1) a longitudinal dataset collected over time and (2) a cross sectional dataset collected over subjects but on the same time. This two datasets pose diverse challenges in that they are susceptible to skewness and are correlated. A

further challenge on the longitudinal data is the individual risk changing over time as measurements apart are always different. This leads us to an interesting challenge of whether the changing in risk profile is important.

This has motivated us to consider different settings to solve health related problems and add knowledge to the large body of literature. Our first approach was to investigate the applicability in binomial data when asymmetry was violated and more particularly estimating the skewness parameter.

In this work, we clearly merged the ideas of [Burr, 1942], [Nagler, 1994] and [McDaniel et al., 2013] together to provide an improved and scientifically supported algorithm that determines the skewness parameter from the GLM framework which we finally fit in the GEE model.

We then tested the performance of the skewness parameter proposed under the GEE by considering a conventional logit (CL) and the skewed logit(SL). The results show that the SL performed better than the CL in bringing out relationship of BV with time as supported by the literature.

The contribution by [Swan, 2006] in addressing the correlation under the tweedie distribution in the GEE framework, motivated us to consider a different type of setting. While he considered longitudinal data on rainfall with zero months when there was no rain recorded, we adopted a different approach by considering clustered data on distance to access inpatient care and cost incurred during outpatient care with zero distance for patients within facility reach and zero cost for the households that did not spend money on outpatient care.

[[Swan, 2006](#)] work adapted the AR(1) correlation structure to find the best predictors for the rainfall, while considering the QICu [[Hardin, 2013](#)]. One of his assumption was that the correlation was from the same subject (same site for rainfall data collection) over time. The AR (1) are good candidates for modelling time delay correlation as they assume closest time points have higher correlation than data collected further apart for the same subject.

We proposed a different approach by considering clustering within a cluster thus our source of correlation was associations within the same (county). Our modeling approach is supported by the fact that people who live in the same cluster are more likely to share similar characteristics such as hospitals, roads and leadership.

Neglecting the correlation would jeopardize our attempts to have better parameter estimates with minimized standard errors in trying to estimate the best predictors for cost and distance. We therefore considered the tweedie distributions under independence and exchangeable correlations to find the best predictors for cost and distance.

A quick attraction to the tweedie under the GEE setting, is in its unique characteristic to accommodate the discrete mass at zero, relax correlation and account for the right skewness in the data. Apart from estimating the best predictors for cost and distance, our work contributed to the statistical methodologies regarding modeling non-normal data and new approaches in estimating the scale parameter as was discussed by [[Hardin, 2013](#)].

GEE are applicable to both normal and non-normal responses, however, if data is non normal and the assumption is violated, then the parameter estimates could be wrong or

underestimated. Our work is therefore very relevant since a lot of biological data is usually non-normal thus providing a new arena of modeling such.

Recommendation

This work further asserts that researchers should have a keen look at the data before modeling, and allow the data to speak for itself rather than forcing conventional ways to model. It further recommends policies targeting the neglected diseases that cause morbidities and mortalities among infants and mothers.

Policies targeting an increase in the number of health facilities and qualified personnel together with proper equipping of public hospitals are encouraged to minimize the long-distance people travel to seek for inpatient care. Alternative policies that would lead to abolishing the cost associated with outpatient care, will also encourage more people to seek the same without the worry of getting poorer.

Further Research

This thesis has investigated several issues related to non-normality under GEE setting. The first two objectives we investigated skewness characteristics in binomial data and selected the model based on its strength to detect a BV time interaction. The conclusions for this approach follow the [[Lipsitz et al., 2019](#)] proposal on selecting models that

can show an association supported by the literature. This approach may not be practical for all the models and therefore other methods of getting the best model would be recommended.

The Quasi Information Criteria (QIC) and the Quasi Information Criteria under independent assumptions QICu proposed by [Pan, 2001] have been found to be great candidates for model selection, and therefore should be modified in such a way that they can be useful in many unique scenarios like we had.

The main problem with our model was the fact that it contained the same number of model covariates, and the QIC and QICu are based on penalizing complex models. Therefore, if we adopt the criteria, we will end up with the same value for the two models and thus can't be compared.

In our other two objectives, we investigated independence and exchangeable characteristic for non-normal responses under the GEE framework, since the AR(1) structure had already been investigated by [Swan, 2006] in modelling rainfall in Australia.

This thesis recommends consideration of other structures in analysis. Some good research problems that future researches can consider include;

- Scenarios where data are balanced but have a non-normal characteristic. Literature supports an unstructured correlation in such scenarios as it is not under any influence, but fits the real correlations. Under such a setting, a good research problem would be, proposed methodological approach for balanced non-normal data using the GEE. In such a scenario, the researchers can modify the R code in this thesis that we adopted for other structures.

- *Proposed hybrid correlation structure under a tweedie distribution* could be a potential topic to investigate. Take a case scenario where public hospitals in counties could be correlated by the fact that they are run by the same body, thus are more likely to receive similar services. While patients who visit these hospitals are only correlated at the level of facility they visit, they have a different correlation from those who visit a different facility (in terms of distance). It would be interesting to see how such a scenario is modelled out
- Modification of QICu to better select the best covariates. We know the QICu criteria by [Pan, 2001] tends to penalize models depending on the number of covariates. Researchers need to look at the penalty imposed by QICu to better select models. A case example is in our table 4.14 and 4.16, we had difficulties in choosing the most appropriate model because they were competing in that some models have low QICu but also low R^2 values, which is statistically true but contradicting. A good research problem would be, *Striking a Balance in model selection when QICu and R^2 are contradicting*

Bibliography

- Affi, A. A., Kotlerman, J. B., Ettner, S. L., and Cowan, M. (2007). Methods for improving regression analysis for skewed continuous or counted responses. *Annual review of public health*, 28:95–111.
- Agarwal, S. K. and Kalla, S. L. (1996). A generalized gamma distribution and its application in reliability. *null*, 25(1):201–210.
- Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In Parzen, E., Tanabe, K., and Kitagawa, G., editors, *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer New York, New York, NY.
- Alcaide, M. L., Chisembele, M., Malupande, E., Arheart, K., Fischl, M., and Jones, D. L. (2015). A cross-sectional study of bacterial vaginosis, intravaginal practices and hiv genital shedding; implications for hiv transmission and women’s health. *BMJ open*, 5:e009036.
- Alcendor, D. J. (2016). Evaluation of health disparity in bacterial vaginosis and the implications for hiv-1 acquisition in african american women. *American journal of reproductive immunology (New York, N.Y. : 1989)*, 76:99–107.
- Allin, S., Masseria, C., and Mossialos, E. (2009). Measuring socioeconomic differences in use of health care services by wealth versus by income. *American journal of public health*, 99(19150899):1849–1855.
- Anigilaje, E. A. (2018). Management of diarrhoeal dehydration in childhood: A review for clinicians in developing countries. *Frontiers in pediatrics*, 6(29527518):28–28.
- Atashili, J., Poole, C., Ndumbe, P. M., Adimora, A. A., and Smith, J. S. (2008). Bacterial vaginosis and hiv acquisition: a meta-analysis of published studies. *AIDS (London, England)*, 22(18614873):1493–1501.
- Awiti, J. O. (2014). Poverty and health care demand in kenya. *BMC Health Services Research*, 14(1):560.
- Awoyemi, T. T., Obayelu, O. A., and Opaluwa, H. I. (2011). Effect of distance on utilization of health care services in rural kogi state, nigeria. *Journal of Human Ecology*, 35(1):1–9.
- Bank, W. (2015). Kenya among the fastest growing economies in africa. resreport, World Bank.

- Barasa, E., Nguhiu, P., and McIntyre, D. (2018a). Measuring progress towards sustainable development goal 3.8 on universal health coverage in kenya. *BMJ Global Health*, 3(3).
- Barasa, E., Rogo, K., Mwaura, N., and Chuma, J. (2018b). Kenya national hospital insurance fund reforms: Implications and lessons for universal health coverage. *Health System and Reforms*, 4(4):346–361.
- Barasa, E. W., Maina, T., and Ravishankar, N. (2017). Assessing the impoverishing effects, and factors associated with the incidence of catastrophic health care payments in kenya. *International Journal for Equity in Health*, 16(1):31.
- Bazán, J. L., Bolfarine, H., and Branco, M. D. (2010). A framework for skew-probit links in binary regression. *Communications in Statistics - Theory and Methods*, 39(4):678–697.
- Benahmed, N., San Miguel, L., Devos, C., Fairon, N., and Christiaens, W. (2017). Vaginal delivery: how does early hospital discharge affect mother and child outcomes? a systematic literature review. *BMC pregnancy and childbirth*, 17:289.
- Berger, S. G., de Pee, S., Bloem, M. W., Halati, S., and Semba, R. D. (2007). Malnutrition and morbidity are higher in children who are missed by periodic vitamin a capsule distribution for child survival in rural indonesia. *The Journal of Nutrition*, 137(5):1328–1333.
- Biswas, R. K. and Kabir, E. (2017). Influence of distance between residence and health facilities on non-communicable diseases: An assessment over hypertension and diabetes in bangladesh. *PLoS one*, 12(28545074):e0177027–e0177027.
- Bono, R., Blanca, M. J., Arnau, J., and Gómez-Benito, J. (2017). Non-normal distributions commonly used in health, education, and social sciences: A systematic review. *Frontiers in psychology*, 8(28959227):1602–1602.
- Brocklehurst, P., Gordon, A., Heatley, E., and Milan, S. J. (2013). Antibiotics for treating bacterial vaginosis in pregnancy. *The Cochrane database of systematic reviews*, page CD000262.
- Bump, R. C. and Buesching, W. J. r. (1988). Bacterial vaginosis in virginal and sexually active adolescent females: evidence against exclusive sexual transmission. *American journal of obstetrics and gynecology*, 158:935–9.
- Burns, D. N., Tuomala, R., Chang, B. H., Hershow, R., Minkoff, H., Rodriguez, E., Zorrilla, C., Hammill, H., and Regan, J. (1997). Vaginal colonization or infection with candida albicans in human immunodeficiency virus-infected women during pregnancy and during the postpartum period. women and infants transmission study group. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 24:201–10.
- Burr, I. W. (1942). Cumulative frequency functions. *Ann. Math. Statist.*, 13(2):215–232.

- Carey, J. C., Klebanoff, M. A., Hauth, J. C., Hillier, S. L., Thom, E. A., Ernest, J., Heine, R. P., Nugent, R. P., Fischer, M. L., Leveno, K. J., Wapner, R., Varner, M., Trout, W., Moawad, A., Sibai, B. M., Miodovnik, M., Dombrowski, M., O'Sullivan, M. J., VanDorsten, J. P., Langer, O., and Roberts, J. (2000). Metronidazole to prevent preterm delivery in pregnant women with asymptomatic bacterial vaginosis. *New England Journal of Medicine*, 342(8):534–540. PMID: 10684911.
- Caron, R., Sinha, D., Dey, D. K., and Polpo, A. (2018). Categorical data analysis using a skewed weibull regression model. *Entropy*, 20(3).
- Castellares, F., Santos, M. A. C., Montenegro, L. C., and Cordeiro, G. M. (2015). A gamma-generated logistic distribution: Properties and inference. *American Journal of Mathematical and Management Sciences*, 34(1):14–39.
- Chaim, W., Mazor, M., and Leiberman, J. R. (1997). The relationship between bacterial vaginosis and preterm birth. a review. *Archives of Gynecology and Obstetrics*, 259(2):51–58.
- Chang, A. Y., Riumallo-Herl, C., Salomon, J. A., Resch, S. C., Brenzel, L., and Verguet, S. (2018). Estimating the distribution of morbidity and mortality of childhood diarrhea, measles, and pneumonia by wealth group in low- and middle-income countries. *BMC Medicine*, 16(1):102.
- Chang, Y. C. (2000). Residuals analysis of the generalized linear models for longitudinal data. *Statistics in medicine*, 19:1277–93.
- Chehoud, C., Stieh, D. J., Bailey, A. G., Laughlin, A. L., Allen, S. A., McCotter, K. L., Sherrill-Mix, S. A., Hope, T. J., and Bushman, F. D. (2017). Associations of the vaginal microbiota with hiv infection, bacterial vaginosis, and demographic factors. *AIDS (London, England)*, 31(28121709):895–904.
- Chen, M.-H., Dey, D. K., and Shao, Q.-M. (1999). A new skewed link model for dichotomous quantal response data. *Journal of the American Statistical Association*, 94(448):1172–1186.
- Chen, M.-H., Dey, D. K., and Shao, Q.-M. (2001). Bayesian analysis of binary data using skewed logit models. *Calcutta Statistical Association Bulletin*, 51(1-2):11–30.
- Chuma, J. and Maina, T. (2012). Catastrophic health care spending and impoverishment in kenya. *BMC Health Services Research*, 12(1):413.
- Coelho, R., Infante, P., and Santos, M. N. (2013). Application of generalized linear models and generalized estimation equations to model at-haulback mortality of blue sharks captured in a pelagic longline fishery in the atlantic ocean. *Fisheries Research*, 145:66 – 75.
- Cohen, C. R., Lingappa, J. R., Baeten, J. M., Ngayo, M. O., Spiegel, C. A., Hong, T., Donnell, D., Celum, C., Kapiga, S., Delany, S., and Bukusi, E. A. (2012). Bacterial vaginosis associated with increased risk of female-to-male hiv-1 transmission: A prospective cohort analysis among african couples. *PLOS Medicine*, 9(6):1–9.

- Cook, R. J. (2000). The theory of dispersion models. bent jørgensen, chapman and hall, 1997. no. of pages: 237. price: £39.95. isbn 0-412-99718-8. *Statistics in Medicine*, 19(14):1952–1953.
- Dingens, A. S., Fairfortune, T. S., Reed, S., and Mitchell, C. (2016). Bacterial vaginosis and adverse outcomes among full-term infants: a cohort study. *BMC pregnancy and childbirth*, 16(27658456):278–278.
- Dunn, P. K. (2017). *Tweedie: Evaluation of Tweedie Exponential Family Models*. R package version 2.3.0.
- Dunn, P. K. and Smyth, G. K. (2005). Series evaluation of tweedie exponential dispersion model densities. *Statistics and Computing*, 15(4):267–280.
- Dunn, P. K. and Smyth, G. K. (2008). Evaluation of tweedie exponential dispersion models using fourier inversion. *Statistics and Computing*, 18:73–86.
- Ensor, T. and Cooper, S. (2004). Overcoming barriers to health service access: influencing the demand side. *Health Policy and Planning*, 19(2):69–79.
- Escamilla, V., Calhoun, L., Winston, J., and Speizer, I. S. (2018). The role of distance and quality on facility selection for maternal and child health services in urban kenya. *Journal of Urban Health*, 95(1):1–12.
- Evans, W. N., Garthwaite, C., and Wei, H. (2008). The impact of early discharge laws on the health of newborns. *Journal of health economics*, 27:843–870.
- Faddy, M., Graves, N., and Pettitt, A. (2009). Modeling length of stay in hospital and other right skewed data: Comparison of phase-type, gamma and log-normal distributions. *Value in Health*, 12(2):309 – 314.
- Farquhar, C., Mbori-Ngacha, D., Overbaugh, J., Wamalwa, D., Harris, J., Bosire, R., and John-Stewart, G. (2010). Illness during pregnancy and bacterial vaginosis are associated with in-utero hiv-1 transmission. *AIDS (London, England)*, 24(19952542):153–155.
- Filmer, D. and Pritchett, L. H. (2001). Estimating wealth effects without expenditure data—or tears: an application to educational enrollments in states of india. *Demography*, 38:115–32.
- Finberg, L. (2002). Dehydration in infancy and childhood. *Pediatrics in Review*, 23(8):277–282.
- Fouda, G. G., Martinez, D. R., Swamy, G. K., and Permar, S. R. (2018). The impact of igg transplacental transfer on early life immunity. *ImmunoHorizons*, 2(29457151):14–25.
- Freedman, D. A. (2006). On the so-called ”huber sandwich estimator” and ”robust standard errors”. *The American Statistician*, 60(4):299–302.
- Freitas, A. C., Chaban, B., Bocking, A., Rocco, M., Yang, S., Hill, J. E., Money, D. M., and Group, V. O. G. U. E. R. (2017). The vaginal microbiome of pregnant women is less rich and diverse, with lower prevalence of mollicutes, compared to non-pregnant women. *Scientific reports*, 7(28835692):9212–9212.

- French, A. L., Adeyemi, O. M., Agniel, D. M., Evans, C. T., Yin, M. T., Anastos, K., and Cohen, M. H. (2011). The association of hiv status with bacterial vaginosis and vitamin d in the united states. *Journal of women's health (2002)*, 20(21875343):1497–1503.
- Gabrysch, S., Cousens, S., Cox, J., and Campbell, O. M. R. (2011). The influence of distance and level of care on delivery place in rural zambia: A study of linked national data in a geographic information system. *PLOS Medicine*, 8(1):e1000394.
- García-Basteiro, A. L., Quintó, L., Macete, E., Bardají, A., González, R., Nhacolo, A., Sigauque, B., Saco, C., Rupérez, M., Sicuri, E., Bassat, Q., Sevene, E., and Menéndez, C. (2017). Infant mortality and morbidity associated with preterm and small-for-gestational-age births in southern mozambique: A retrospective cohort study. *PLOS ONE*, 12(2):1–14.
- Gibson, J. P. (1959). Control of persistent vomiting in infants and children. *Pediatrics*, 23(3):578–581.
- Gilchrist, R. and Drinkwater, D. (2000). The use of the tweedie distribution in statistical modelling. In Bethlehem, J. G. and van der Heijden, P. G. M., editors, *COMPSTAT*, pages 313–318, Heidelberg. Physica-Verlag HD.
- Giner, G. and Smyth, G. K. (2016). statmod: probability calculations for the inverse gaussian distribution. *R Journal*, 8(1):339–351.
- Girma, F., Jira, C., and Girma, B. (2011). Health services utilization and associated factors in jimma zone, south west ethiopia. *Ethiopian journal of health sciences*, 21:85–94.
- GOK (2014). 2013 kenya household health expenditure and utilisation survey. Technical report, Ministry of Health.
- Goleř, I. (2014). Symmetric and asymmetric binary choice models for corporate bankruptcy. *Procedia - Social and Behavioral Sciences*, 124:282 – 291. Challenges and Innovations in Management and Leadership.
- Group, O. B. (2020). New health care initiatives in kenya to increase access and quality. Technical report, Oxford Business Group.
- Guaschino, S., De Seta, F., Piccoli, M., Maso, G., and Alberico, S. (2006). Aetiology of preterm labour: bacterial vaginosis. *BJOG : an international journal of obstetrics and gynaecology*, 113 Suppl 3:46–51.
- Haggerty, C. L., Hillier, S. L., Bass, D. C., Ness, R. B., Evaluation, P. I. D., and Investigators, C. H. P. S. (2004). Bacterial vaginosis and anaerobic bacteria are associated with endometritis. *Clinical Infectious Diseases*, 39(7):990–995.
- Hardin, J. W. (2013). *Generalized estimating equations*. Hardin, Hardin,. CRC Press, Boca Raton, Fla.
- Hasan, M. M. and Dunn, P. K. (2012). Understanding the effect of climatology on monthly rainfall amounts in australia using tweedie glms. *International Journal of Climatology*, 32(7):1006–1017.

- Hay, J., Walker, A., Sanchez, K., and Thompson, K. (2019). Abstract social categories facilitate access to socially skewed words. *PLOS ONE*, 14(2):e0210793.
- Hernandez, J. B. R. and Kim, P. Y. (2020). Epidemiology morbidity and mortality.
- Hillier, S. L., Nugent, R. P., Eschenbach, D. A., Krohn, M. A., Gibbs, R. S., Martin, D. H., Cotch, M. F., Edelman, R., Pastorek, J. G., Rao, A. V., McNellis, D., Regan, J. A., Carey, J. C., and Klebanoff, M. A. (2018). Association between bacterial vaginosis and preterm delivery of a low-birth-weight infant. *N Engl J Med*, 333(26):1737–1742.
- Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101.
- Ibrahim, J. G., Chu, H., and Chen, L. M. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 28(20439643):2796–2801.
- Ilinca, S., Di Giorgio, L., Salari, P., and Chuma, J. (2019). Socio-economic inequality and inequity in use of health care services in kenya: evidence from the fourth kenya household health expenditure and utilization survey. *International Journal for Equity in Health*, 18(1):196.
- Isik, G., Demirezen, s., Dönmez, H. G., and Beksac, M. S. (2016). Bacterial vaginosis in association with spontaneous abortion and recurrent pregnancy losses. *Journal of cytology*, 33(27756985):135–140.
- Jallow, S., Cutland, C. L., Masbou, A. K., Adrian, P., and Madhi, S. A. (2017). Maternal hiv infection associated with reduced transplacental transfer of measles antibodies and increased susceptibility to disease. *Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology*, 94:50–56.
- Jamieson, D. J., Duerr, A., Klein, R. S., Paramsothy, P., Brown, W., Cu-Uvin, S., Rompalo, A., and Sobel, J. (2001). Longitudinal analysis of bacterial vaginosis: findings from the hiv epidemiology research study. *Obstetrics and gynecology*, 98:656–63.
- Jing, R., Xu, T., Lai, X., Mahmoudi, E., and Fang, H. (2020). Technical efficiency of public and private hospitals in beijing, china: A comparative study. *International Journal of Environmental Research and Public Health*, 17(1).
- Jones, E., Taylor, B., Rudge, G., MacArthur, C., Jyothish, D., Simkiss, D., and Cummins, C. (2018). Hospitalisation after birth of infants: cross sectional analysis of potentially avoidable admissions across england using hospital episode statistics. *BMC Pediatrics*, 18(1):390.
- Kadobera, D., Sartorius, B., Masanja, H., Mathew, A., and Waiswa, P. (2012). The effect of distance to formal health facility on childhood mortality in rural tanzania, 2005-2007. *Global health action*, 5:1–9.
- Kamga, Y. M., Ngunde, J. P., and Akoachere, J.-F. K. T. (2019). Prevalence of bacterial vaginosis and associated risk factors in pregnant women receiving antenatal care at the kumba health district (khd), cameroon. *BMC Pregnancy and Childbirth*, 19(1):166.

- Karra, M., Fink, G., and Canning, D. (2017). Facility distance and child mortality: a multi-country study of health facility access, service utilization, and child health outcomes. *Int J Epidemiol*, 46(3):817–826.
- Kato, R. and Okada, M. (2019). Can financial support reduce suicide mortality rates? *International Journal of Environmental Research and Public Health*, 16(23).
- Keats, E. C., Ngugi, A., Macharia, W., Akseer, N., Khaemba, E. N., Bhatti, Z., Rizvi, A., Tole, J., and Bhutta, Z. A. (2019). Progress and priorities for reproductive, maternal, newborn, and child health in kenya: a countdown to 2015 country case study. *The Lancet Global Health*, 5(8):e782–e795.
- Kellerman, S. E., Ahmed, S., Feeley-Summerl, T., Jay, J., Kim, M., Phelps, B. R., Sugandhi, N., Schouten, E., Tolle, M., Tsiouris, F., of the Interagency Task Team on the Prevention, C. S. W. G., Treatment of HIV infection in Pregnant Women, M., and Children (2013). Beyond prevention of mother-to-child transmission: keeping hiv-exposed and hiv-positive children healthy and alive. *AIDS (London, England)*, 27 Suppl 2(24361632):S225–S233.
- Kelly, C., Hulme, C., Farragher, T., and Clarke, G. (2016). Are differences in travel time or distance to healthcare for adults in global north countries associated with an impact on health outcomes? a systematic review. *BMJ open*, 6(27884848):e013059–e013059.
- Kevany, S., Murima, O., Singh, B., Hlubinka, D., Kulich, M., Morin, S. F., and Sweat, M. (2012). Socio-economic status and health care utilization in rural zimbabwe: findings from project accept (hptn 043). *Journal of Public Health in Africa*, 3(1):e13.
- Khalil, I. A., Troeger, C., Rao, P. C., Blacker, B. F., Brown, A., Brewer, T. G., Colombara, D. V., De Hostos, E. L., Engmann, C., Guerrant, R. L., Haque, R., Houpt, E. R., Kang, G., Korpe, P. S., Kotloff, K. L., Lima, A. A. M., Petri, William A, J., Platts-Mills, J. A., Shoultz, D. A., Forouzanfar, M. H., Hay, S. I., Reiner, Robert C, J., and Mokdad, A. H. (2019). Morbidity, mortality, and long-term consequences associated with diarrhoea from *cryptosporidium* infection in children younger than 5 years: a meta-analysis study. *The Lancet Global Health*, 6(7):e758–e768.
- Kimani, D. N. (2014). *Out-of-pocket health expenditures and household poverty: evidence from kenya*. PhD thesis, University of Nairobi.
- Kimathi, L. (2017). Challenges of the devolved health sector in kenya: Teething problems or systemic contradictions? *Africa Development / Afrique et Développement*, 42(1):55–77.
- Kinney, M. V., Kerber, K. J., Black, R. E., Cohen, B., Nkrumah, F., Coovadia, H., Nampala, P. M., Lawn, J. E., on behalf of the Science in Action: Saving the lives of Africa’s mothers, n., and children working group (2010). Sub-saharan africa’s mothers, newborns, and children: Where and why do they die? *PLOS Medicine*, 7(6):e1000294.
- Kinshella, M.-L. W., Walker, C. R., Hiwa, T., Vidler, M., Nyondo-Mipando, A. L., Dube, Q., Goldfarb, D. M., and Kawaza, K. (2020). Barriers and facilitators to implementing bubble cpap to improve neonatal health in sub-saharan africa: a systematic review. *Public Health Reviews*, 41(1):6.

- KNBS (2015). Kenya demographic and health survey 2014. Technical report, Kenya National Bureau of Statistics, Rockville, MD, USA.
- Kukla, M., McKay, N., Rheingans, R., Harman, J., Schumacher, J., Kotloff, K. L., Levine, M. M., Breiman, R., Farag, T., Walker, D., Nasrin, D., Omoro, R., O'Reilly, C., and Mintz, E. (2017). The effect of costs on Kenyan households' demand for medical care: why time and distance matter. *Health Policy and Planning*, 32(10):1397–1406.
- Kurz, C. F. (2017). Tweedie distributions for fitting semicontinuous health care utilization cost data. *BMC Medical Research Methodology*, 17(1):171.
- Ladner, J., Besson, M.-H., Rodrigues, M., Sams, K., Audureau, E., and Saba, J. (2013). Prevention of mother-to-child hiv transmission in resource-limited settings: assessment of 99 viramune donation programmes in 34 countries, 2000-2011. *BMC public health*, 13(23672811):470–470.
- Laird, N. M. (1988). Missing data in longitudinal studies. *Statistics in Medicine*, 7(2):305–315.
- Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.
- Lassi, Z. S., Majeed, A., Rashid, S., Yakoob, M. Y., and Bhutta, Z. A. (2013). The interconnections between maternal and newborn health – evidence and implications for policy. *The Journal of Maternal-Fetal & Neonatal Medicine*, 26(sup1):3–53.
- Lean, R. E., Rogers, C. E., Paul, R. A., and Gerstein, E. D. (2018). Nicu hospitalization: Long-term implications on parenting and child behaviors. *Current treatment options in pediatrics*, 4(29881666):49–69.
- Lee, W.-Y. and Shaw, I. (2014). The impact of out-of-pocket payments on health care inequity: The case of national health insurance in south korea. *International Journal of Environmental Research and Public Health*, 11(7):7304–7318.
- Leidman, E., Mwirigi, L. M., Maina-Gathigi, L., Wamae, A., Imbwaga, A. A., and Bilukha, O. O. (2018). Assessment of anthropometric data following investments to ensure quality: Kenya demographic health surveys case study, 2008 to 2009 and 2014. *Food and nutrition bulletin*, 39(3):406–419.
- Lepargneur, J. P. and Rousseau, V. (2002). [protective role of the doederlein flora]. *Journal de gynecologie, obstetrique et biologie de la reproduction*, 31:485–94.
- Li, A., Shi, Y., Yang, X., and Wang, Z. (2019a). Effect of critical illness insurance on household catastrophic health expenditure: The latest evidence from the national health service survey in china. *International Journal of Environmental Research and Public Health*, 16(24).
- Li, L., Jiang, J., Xiang, L., Wang, X., Zeng, L., and Zhong, Z. (2019b). Impact of critical illness insurance on the burden of high-cost rural residents in central china: An interrupted time series study. *International Journal of Environmental Research and Public Health*, 16(19).

- Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.
- Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J. G., Sinha, D., Parzen, M., and Lipshultz, S. (2019). Joint generalized estimating equations for multivariate longitudinal binary outcomes with missing data: an application to acquired immune deficiency syndrome data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1):3–20.
- Liu, H. and Dai, W. (2020). An empirical study on the benefits equity of the medical security policy: the china health and nutrition survey (chns). *International Journal of Environmental Research and Public Health*, 17(4).
- Manandhar, B. and Nandram, B. (2019). Hierarchical bayesian models for continuous and positively skewed data from small areas. *Communications in Statistics - Theory and Methods*, pages 1–19.
- Manikandan, S. (2010). Data transformation. *Journal of pharmacology & pharmacotherapeutics*, 1:126–7.
- Mbau, R., Kabia, E., Honda, A., Hanson, K., and Barasa, E. (2020). Examining purchasing reforms towards universal health coverage by the national hospital insurance fund in kenya. *International Journal for Equity in Health*, 19(1):19.
- Mbori-Ngacha, D., Nduati, R., John, G., and et al (2001). Morbidity and mortality in breastfed and formula-fed infants of hiv-1-infected women: A randomized clinical trial. *JAMA*, 286(19):2413–2420.
- McCullagh, P. (1984). Generalized linear models. *European Journal of Operational Research*, 16(3):285 – 292.
- McDaniel, L. S., Henderson, N. C., and Rathouz, P. J. (2013). Fast pure R implementation of GEE: application of the Matrix package. *The R Journal*, 5:181–187.
- McGregor, J. A. and French, J. I. (2000). Bacterial vaginosis in pregnancy. *Obstetrical & gynecological survey*, 55:S1–19.
- Monebenimp, F., Nga-Essono, D. E., Zoung-Kany Bissek, A.-C., Chelo, D., and Tetanye, E. (2011). Hiv exposure and related newborn morbidity and mortality in the university teaching hospital of yaoundé, cameroon. *The Pan African medical journal*, 8(22121451):43–43.
- Muriithi, M. K. (2013). The determinants of health-seeking behavior in a nairobi slum, kenya. *European Scientific Journal, ESJ*, 9(8).
- Murphy, M. S. (2008). Management of bloody diarrhoea in children in primary care. *BMJ (Clinical research ed.)*, 336:1010–5.
- Mwabu, G. M. (1989). Nonmonetary factors in the household choice of medical facilities. *Economic Development and Cultural Change*, 37(2):383–392.

- Mwaliko, E., Downing, R., O Meara, W., Chelagat, D., Obala, A., Downing, T., Simiyu, C., Odhiambo, D., Ayuo, P., Menya, D., and Khwa-Otsyula, B. (2014). "not too far to walk": the influence of distance on place of delivery in a western kenya health demographic surveillance system. *BMC Health Services Research*, 14(1):212.
- Mwenda, N., Nduati, R., Kosgei, M., and Kerich, G. (2021a). Skewed logit model for analyzing correlated infant morbidity data. *PLOS ONE*, 16(2):1–16.
- Mwenda, N., Nduati, R., Kosgey, M., and Kerich, G. (2021b). Effect of bacterial vaginosis (bv)-hiv-1 co-existence on maternal and infant health: A secondary data analysis. *Frontiers in Pediatrics*, 9:191.
- Nagler, J. (1994). Scobit: An alternative estimator to logit and probit. *American Journal of Political Science*, 38(1):230–255.
- Nduati, R., John, G., Mbori-Ngacha, D., Richardson, B., Overbaugh, J., Mwatha, A., Ndinya-Achola, J., Bwayo, J., Onyango, F. E., Hughes, J., and Kreiss, J. (2000). Effect of breastfeeding and formula feeding on transmission of hiv-1: a randomized clinical trial. *JAMA*, 283:1167–74.
- Nduati, R., Richardson, B. A., John, G., Mbori-Ngacha, D., Mwatha, A., Ndinya-Achola, J., Bwayo, J., Onyango, F. E., and Kreiss, J. (2001). Effect of breastfeeding on mortality among hiv-1 infected women: a randomised trial. *Lancet (London, England)*, 357:1651–5.
- Nesbitt, R. C., Lohela, T. J., Soremekun, S., Vesel, L., Manu, A., Okyere, E., Grundy, C., Amenga-Etego, S., Owusu-Agyei, S., Kirkwood, B. R., and Gabrysch, S. (2016). The influence of distance and quality of care on place of delivery in rural ghana. *Scientific reports*, 6:30291.
- Ngugi, A. K., Agoi, F., Mahoney, M. R., Lakhani, A., Mang'ong'o, D., Nderitu, E., Armstrong, R., and Macfarlane, S. (2017). Utilization of health services in a resource-limited rural area in kenya: Prevalence and associated household-level factors. *PLOS ONE*, 12(2):1–12.
- Nic Carthaigh, N., De Gryse, B., Esmati, A. S., Nizar, B., Van Overloop, C., Fricke, R., Bseiso, J., Baker, C., Decroo, T., and Philips, M. (2014). Patients struggle to access effective health care due to ongoing violence, distance, costs and health service performance in afghanistan. *Int Health*, 7(3):169–175.
- Noor, A. M., Amin, A. A., Gething, P. W., Atkinson, P. M., Hay, S. I., and Snow, R. W. (2006). Modelling distances travelled to government health services in kenya. *Tropical medicine & international health : TM & IH*, 11:188–96.
- Obare, V., Brolan, C. E., and Hill, P. S. (2014). Indicators for universal health coverage: can kenya comply with the proposed post-2015 monitoring recommendations? *International Journal for Equity in Health*, 13(1):123.
- OECD (2020). Employment rate by age group(indicator).

- Okech, T. C. and Lelegwe, S. L. (2015). Analysis of universal health coverage and equity on health care in kenya. *Global journal of health science*, 8(26925910):218–227.
- Pan, W. (2001). Akaike’s information criterion in generalized estimating equations. *Biometrics*, 57(1):120–125.
- Peña-Sanchez, I. (2019). Applying the tweedie model for improved microinsurance pricing. *The Geneva Papers on Risk and Insurance - Issues and Practice*, 44(3):365–381.
- Pezzulo, C., Hornby, G. M., Sorichetta, A., Gaughan, A. E., Linard, C., Bird, T. J., Kerr, D., Lloyd, C. T., and Tatem, A. J. (2017). Sub-national mapping of population pyramids and dependency ratios in africa and asia. *Scientific Data*, 4(1):170089.
- Posel, D. R. (2001). Who are the heads of household, what do they do, and is the concept of headship useful? an analysis of headship in south africa. *null*, 18(5):651–670.
- Prentice, R. L. (1976). A generalization of the probit and logit methods for dose response curves. *Biometrics*, 32:761–8.
- Priestley, C. J., Jones, B. M., Dhar, J., and Goodwin, L. (1997). What is normal vaginal flora? *Genitourinary medicine*, 73(9155551):23–28.
- R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rabarison, K. M., Bish, C. L., Massoudi, M. S., and Giles, W. H. (2015). Economic evaluation enhances public health decision making. *Frontiers in Public Health*, 3:164.
- Rathie, P. N., Silva, P., and Olinto, G. (2016). Applications of skew models using generalized logistic distribution. *Axioms*, 5(2).
- Ravikumara, M. and Bhat, B. V. (1996). Early neonatal mortality in an intramural birth cohort at a tertiary care hospital. *Indian journal of pediatrics*, 63:785–9.
- Rippin, H. L., Hutchinson, J., Greenwood, D. C., Jewell, J., Breda, J. J., Martin, A., Rippin, D. M., Schindler, K., Rust, P., Fagt, S., Matthiessen, J., Nurk, E., Nelis, K., Kukk, M., Tapanainen, H., Valsta, L., Heuer, T., Sarkadi-Nagy, E., Bakacs, M., Tazhibayev, S., Sharmanov, T., Spiroski, I., Beukers, M., van Rossum, C., Ocke, M., Lindroos, A. K., Warensjo Lemming, E., and Cade, J. E. (2020). Inequalities in education and national income are associated with poorer diet: Pooled analysis of individual participant data across 12 european countries. *PLOS ONE*, 15(5):e0232447.
- Rizopoulos, D., Molenberghs, G., and Lesaffre, E. M. E. H. (2017). Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. *Biometrical journal. Biometrische Zeitschrift*, 59:1261–1276.
- Robertson, L., Skelly, C., and Phillips, D. (2019). Making hard choices in local public health spending with a cost-benefit analysis approach. *Frontiers in Public Health*, 7:147.
- Salari, P., Di Giorgio, L., Ilinca, S., and Chuma, J. (2019). The catastrophic and impoverishing effects of out-of-pocket healthcare payments in kenya, 2018. *BMJ Global Health*, 4(6).

- Salzer, H. E. (1969). Chebyshev interpolation and quadrature formulas of very high degree. *Commun. ACM*, 12(5):271.
- Schmid, G., Markowitz, L., Joesoef, R., and Koumans, E. (2000). Bacterial vaginosis and hiv infection. *Sexually Transmitted Infections*, 76(1):3–4.
- Schoeps, A., Gabrysch, S., Niamba, L., Sié, A., and Becher, H. (2011). The effect of distance to health-care facilities on childhood mortality in rural burkina faso. *Am J Epidemiol*, 173(5):492–498.
- Sha, B. E., Zariffard, M. R., Wang, Q. J., Chen, H. Y., Bremer, J., Cohen, M. H., and Spear, G. T. (2005). Female genital-tract hiv load correlates inversely with lactobacillus species but positively with bacterial vaginosis and mycoplasma hominis. *The Journal of infectious diseases*, 191:25–32.
- Shapiro, R. L., Lockman, S., Kim, S., Smeaton, L., Rahkola, J. T., Thior, I., Wester, C., Moffat, C., Arimi, P., Ndase, P., Asmelash, A., Stevens, L., Montano, M., Makhema, J., Essex, M., and Janoff, E. N. (2007). Infant morbidity, mortality, and breast milk immunologic profiles among breast-feeding hiv-infected and hiv-uninfected women in botswana. *The Journal of infectious diseases*, 196:562–9.
- Shiva, F., Sanaei Dashti, A., and Hosseini Khorami, H. (2017). Causes and risk factors of hospitalization among infants less than six months old in tehran. *Archives of Pediatric Infectious Diseases*, 5(3):e33722.
- Shon, C., Lee, T. H., Ndombi, G. O., and Nam, E. W. (2018). A cost-benefit analysis of the official development assistance project on maternal and child health in kwango, dr congo. *International Journal of Environmental Research and Public Health*, 15(7).
- Smyth, G. K. and Jørgensen, B. (2002). Fitting tweedie’s compound poisson model to insurance claims data: Dispersion modelling. *ASTIN Bulletin*, 32(1):143–157.
- Spear, G. T., St John, E., and Zariffard, M. (2007). Bacterial vaginosis and human immunodeficiency virus infection. *AIDS Research and Therapy*, 4(1):25.
- StataCorp (2015). *Stata Statistical Software: Release 14*. College Station, TX: StataCorp LP. College Station, TX, 14 edition.
- Stevenson, D. K., Verter, J., Fanaroff, A. A., Oh, W., Ehrenkranz, R. A., Shankaran, S., Donovan, E. F., Wright, L. L., Lemons, J. A., Tyson, J. E., Korones, S. B., Bauer, C. R., Stoll, B. J., and Papile, L.-A. (2000). Sex differences in outcomes of very low birthweight infants: the newborn male disadvantage. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, 83(3):F182–F185.
- Stock, R. (1983). Distance and the utilization of health facilities in rural nigeria. *Social Science & Medicine*, 17(9):563 – 570.
- Su, S., Dzipire, N. C., Ngare, P., and Odongo, L. (2018). A poisson-gamma model for zero inflated rainfall data. *Journal of Probability and Statistics*, 2018:1012647.

- Sun, G. W., Shook, T. L., and Kay, G. L. (1996). Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *Journal of clinical epidemiology*, 49:907–16.
- Swallow, B., Buckland, S. T., King, R., and Toms, M. P. (2016). Bayesian hierarchical modelling of continuous non-negative longitudinal data with a spike at zero: An application to a study of birds visiting gardens in winter. *Biometrical journal. Biometrische Zeitschrift*, 58:357–71.
- Swan, T. (2006). *Generalized estimating equations when the response variable has a Tweedie distribution: An application for multi-site rainfall modelling*. PhD thesis, The University of Southern Queensland.
- Tay, R. (2016). Comparison of the binary logistic and skewed logistic (scobit) models of injury severity in motor vehicle collisions. *Accident Analysis & Prevention*, 88:52–55.
- Thaddeus, S. and Maine, D. (1994). Too far to walk: maternal mortality in context. *Social science & medicine (1982)*, 38:1091–110.
- Thorsen, P., Vogel, I., Olsen, J., Jeune, B., Westergaard, J. G., Jacobsson, B., and Moller, B. R. (2006). Bacterial vaginosis in early pregnancy is associated with low birth weight and small for gestational age, but not with spontaneous preterm birth: a population-based study on danish women. *The journal of maternal-fetal & neonatal medicine : the official journal of the European Association of Perinatal Medicine, the Federation of Asia and Oceania Perinatal Societies, the International Society of Perinatal Obstetricians*, 19:1–7.
- Tlou, B., Sartorius, B., and Tanser, F. (2018). Investigating risk factors for under-five mortality in an hiv hyper-endemic area of rural south africa, from 2000-2014. *PLOS ONE*, 13(11):1–15.
- Ukah, U. V., Bayrampour, H., Sabr, Y., Razaz, N., Chan, W.-S., Lim, K. I., and Lisonkova, S. (2020). Association between gestational weight gain and severe adverse birth outcomes in washington state, us: A population-based retrospective cohort study, 2004–2013. *PLOS Medicine*, 16(12):1–17.
- Umar, N., Litaker, D., Schaarschmidt, M.-L., Peitsch, W. K., Schmieder, A., and Terris, D. D. (2012). Outcomes associated with matching patients’ treatment preferences to physicians’ recommendations: study methodology. *BMC Health Services Research*, 12(1):1.
- van der Heyden, J. L., van Kuijk, S. M. J., van der Ham, D. P., Notten, K. J. B., Janssen, T., Nijhuis, J. G., Willekes, C., Porath, M., van der Post, J. A., Halbertsma, F., Pajkrt, E., and Mol, B. W. J. (2013). Subsequent pregnancy after preterm prelabor rupture of membranes before 27 weeks’ gestation. *AJP reports*, 3(24147248):113–118.
- Venkatesh, K. K., de Bruyn, G., Marinda, E., Otwombe, K., van Niekerk, R., Urban, M., Triche, E. W., McGarvey, S. T., Lurie, M. N., and Gray, G. E. (2011). Morbidity and mortality among infants born to hiv-infected women in south africa: implications for child health in resource-limited settings. *Journal of tropical pediatrics*, 57(20601692):109–119.

- Verma, I. C. and Kumar, S. (1968). Causes of morbidity in children attending a primary health centre. *The Indian Journal of Pediatrics*, 35(12):543–549.
- Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the gauss–newton method. *Biometrika*, 61(3):439–447.
- Wickham, H., François, R., Henry, L., and Müller, K. (2020). *dplyr: A Grammar of Data Manipulation*. R package version 1.0.2.
- Wright, E. M. (1933). On the coefficients of power series having exponential singularities. *Journal of the London Mathematical Society*, s1-8(1):71–79.
- Wright, N. D., Symmonds, M., Morris, L. S., and Dolan, R. J. (2013). Dissociable influences of skewness and valence on economic choice and neural activity. *PLOS ONE*, 8(12):e83454.
- Zeileis, A. and Windberger, T. (2018). *glogis: Fitting and Testing Generalized Logistic Distributions*. R package version 1.0-1.
- Zhang, J. and Timmermans, H. (2019). Scobit-based panel analysis of multitasking behavior of public transport users. *Transportation Research Record*, 2157(1):46–53.
- Zhang, Z. (2016). Variable selection with stepwise and best subset approaches. *Annals of translational medicine*, 4:136.

Appendix A: Obtaining the disturbance term α for use in section 4.3

Notations and statistical methods

Consider n independent subjects observed at a specified time t . Given the number of subjects in our data, let, $i = 1, \dots, n$ where $n = 327$ represents the total number of subjects and we let the i^{th} infant be observed n_i times, where $j = 1, \dots, 6$. A subject could be observed at a common set of time $t = 1, \dots, m$ up to a maximum of 6 times. The methods can also be applied to unequally spaced time $t_1 < \dots < t_m$ points (see Hardie and Hilbe [[Hardin, 2013](#)]. pg 75).

Let $x_i = (x_{i1}, \dots, x_{in_j})^T$ represent an $n_i \times p$ matrix-vector of covariates and let y_{it} be an $n_i \times 1$ vector of responses. This study assumes that our response variable, Y_{ij} has a Bernoulli distribution, i.e.,

$$y_{ij} \mid \rho_{ij} \sim \text{Bern}(\rho_{ij}) \tag{1}$$

with unknown

$$E(y_{ij}) = \rho_{ij} \quad (2)$$

The outcome at time t which is morbidity can be represented as $Y_i = 1$ if Yes for morbidity and 0 otherwise This dependent variable is related to the covariates through a link function given by

$$g(\rho_{ij}) = x'_{ij}\beta \quad (3)$$

, where $g(\cdot)$ is a logit link function, β is a p - dimensional vector of regression coefficients, and x_{ij} is an n - dimensional vector of covariates.

To model the marginal and subject specific probability of this type of response, authors have suggested we parametrize using a probit link $\Phi^{-1}\mu$ or a logit link

$$\ln\left(\frac{\mu}{1-\mu}\right) \quad (4)$$

Subject specific analyses are important since their interpretation is at a lower level, but are complex to handle when there is a large number of respondents. A probit link function given by

$$\Phi\left(\frac{X_{it}\beta^{ss}}{\sqrt{1+\sigma_v^2}}\right) \quad (5)$$

is a good candidate for this type, but has computation complexity for modeling higher order associations, while the logit link function given by

$$\Phi\left(\frac{X_{it}\beta^{ss}}{\sqrt{1+Q\sigma_v^2}}\right) \quad (6)$$

is easy to implement but has no closed form. The constant is expressed by

$$Q = 16\sqrt{3/15\pi} \quad (7)$$

This poses a challenge in the interpret ability and practical applicability of the model.

Furthermore, the link functions are widely applicable in GLMs, which require that the full likelihood be specified. A major drawback to this approach for repeated measures is that an increase in the number of the measures results in an exponential increase in the number of parameters to be specified in the model and estimated. Both the logit and probit have conditional probability distributions, which are maximum at 0 such that P_i for $i \in (0, 1)$ is 0.5 and thus has a fixed symmetry at 0.5. However, symmetry may not be realistic to all Bernoulli or continuous responses as demonstrated by the works of different researchers in different fields such as political science by [Nagler, 1994], social and behavioral sciences by [Golet, 2014], and fisheries by [Coelho et al., 2013]. In these entire analyses, the researchers obtained better results by ignoring normality in the response.

The methods used by Nagler followed an asymmetric logistic distribution and his results hold for models without repeated measures. Therefore, there was an assumption of independence among the responses. The restriction to models with independence is unappealing to models in a longitudinal set up where there is correlation within the subject measurements with time and interaction of covariates with time is of essence. The research conducted by [Coelho et al., 2013] was based on the GEE framework but their response was continuous.

As far as we are concerned, we have not come across any work on asymmetric binary under the GEE framework. We constructed a flexible link function that can accommodate both symmetric and asymmetric binary responses. Consider the model in which the response Y_i relates to the latent Y_i^* as follows:

$$Y_i = \begin{cases} 1, & \text{if } Y_i^* > 0 \\ 0, & \text{Otherwise} \end{cases} \quad (8)$$

The probability density function (PDF) of subject i falling in the category of morbidity incidence as given by Nagler is

$$P_i = \Pr[X_i\beta + \mu_i > 0] \quad (9)$$

and the cumulative distribution function (CDF) is given by

$$P_i = 1 - F(-X_i\beta) \quad (10)$$

such that Y_i can be expressed as

$$P_i(Y_0 = 1) = F(-X_i\beta) \quad (11)$$

The marginal effect on P_i for a change in X_m is expressed as

$$\frac{\partial P_i}{\partial(X_m)} = \frac{\partial[1 - F(-X_i\beta)]}{\partial(X_m)} = f(-X_i\beta)\beta_m \quad (12)$$

We estimate the probability P_i for which the sensitivity $\partial P_i / \partial(X_m)$ is the maximum. To relax the strong conditional probability on a binary response, accommodate the heterogeneity of repeated measures on the subjects, and put up for interaction time effects in the selected covariates, we employ the Burr type 10 distribution in the logit link under the GEE. This caters for the disturbance introduced in the logit during this process.

Therefore, as proposed by Nagler, we have a more flexible predictor that can handle both symmetric and asymmetric responses in the binary variable. However, to modify the link function, the logit link is usually preferred as it is easier to modify and can easily be generalized to imitate or mimic the Burr type 10 distribution proposed by Irving Burr in 1942 [Burr \[1942\]](#) which is a desired characteristic in this work. We introduce another parameter ϕ that will be referred to as the skewness parameter, and will be used to modify our response curve. This variation implies that the maximum is no longer restricted to $P = 0.5$.

The disturbance term estimated as $\hat{\alpha}$ is independent of time and therefore its GLM estimate is assumed to be unbiased for a true α . This is achieved through an iterative weighted least square method put forth by [McCullagh, 1984](#). An advantage of this model extension is that the disturbance term is assumed to be a constant. Therefore, our model can still be easily generalized to conform to the exponential dispersion model (EDM) which can then be easily adopted in the GEE framework. [\[Burr, 1942\]](#) developed the Burr type 10 distribution given a random variable X with a cumulative distribution $F(x)$ with its CDF distribution defined as

$$F(x) = \begin{cases} 1 - \frac{1}{(1+x^c)^k}, & \text{if } x \geq 0 \\ 0, & \text{Otherwise} \end{cases} \quad (13)$$

with PDF given by

$$F'(x) = f(x) = 1 - \frac{kcx^{c-1}}{(1+x^c)^{k+1}} \quad (14)$$

John Nagler recently proposed a new family of distributions called the skewed logit in which he modified the *Burr type 10 CDF* to mimic the logit such that the characteristic 'S'-shaped curve of the sigmoid function

$$S(x) = \frac{1}{1 + e^{-x}} \quad (15)$$

is retained to accommodate a binary response.

This was achieved by adding a constant parameter

$$F(k; \phi) = \frac{1}{(1 + e^{-k})^\phi} \quad (16)$$

, for $\phi > 0$, which is non-zero, non-negative and continuous, to the logit distribution function. This makes the estimator more flexible in modeling real binary data, since the logit is now nested in the scobit such that when the parameter is 1, then the proposed estimator conforms to the logistic distribution (see Figure 3.2 for various values of the

parameter). As can be seen, the sigmoid slope takes on its maximum values at different probability levels depending on the parameter of choice.

There is a myriad of approaches to estimating skewness particularly under the GLM framework in which independence is assumed and the conventional way is to model binary 'conditional independence models'. We propose a new way of modeling binary data referred to as the 'unconditional dependence model' a few methods of which exist.

This paper proposes a modification of the link function in the GEE proposed by Liang and Zeger to handle asymmetry in data with dependence. The skewed logit proposed under the GEE framework has a multiplier with a constant, meaning it can still be expressed as an EDM. This property makes it very easy to integrate in the GEE as the mean-variance relationship can easily be estimated.

This property also implies that we do not have a difficult task in specifying a full likelihood (in which we arrive at wrong conclusions when we specify a wrong one) but we could flexibly select any correlation structure and still obtain plausible results and reduction in the margin of error and bias. Since we are dealing with a binary response calculated as a score from a continuous response, the common approach would be to assume the logit model given by

$$\log \left(\frac{\mu}{1 - \mu} = \beta X_i^T \in \mathfrak{R} \right) \quad (17)$$

where X_i 's are the model covariates that include the weight, mother's BV status, HIV status of the infants, and feeding status in our data and the β values are the coefficients to be estimated. The logit assumes that the probability of success or failure is the same,

maintaining symmetry assumptions and the maximum of the logit is achieved at $p = 0.5$. In this work, we aim to consider a response that violates the symmetry assumption, but still within the same framework. Let $k(\cdot)$ be the link function and $E(Y) = \mu$ such that $k(\mu) = X\beta$. The $k^{(-1)}$ is the logit link for a binary response.

$$K^{-1}(\mu_{ij}) = \eta_{ij} = X_{ij}^T \quad (18)$$

Fitting a GLM to the data to obtain the disturbance parameter

After specifying the parameters, the initial estimate of β and the disturbance term were obtained using the GLM approach using scobit as the link function. However, the β values are just a proxy of association or what researchers refer to as “starting values” and not correct since they assume the presence of independence. The assumption of independence implies that the standard errors could be underestimated or overestimated, because we ignore within-subject dependency.

Assuming that the parameter is a constant and that the scobit and logit are related, then by a direct relationship, a scobit belongs to the EDM defined by the marginal density of the response belonging to the family of exponential distributions. For the repeated Bernoulli response where measurements for subject i are taken repeatedly at time t , the PDF is given by

$$\exp \left\{ \frac{Y_{it}\theta_{it} - b(\theta_{it})}{a\Phi} + c(Y_{it}, \Phi) \right\} \quad (19)$$

where θ is the natural or canonical parameter, $a_i(\phi) > 0$ is the scale parameter, and $c(y, \phi)$ is the normalizing constant to ensure the PDF integrates to one. From basic

principles, it can easily be shown that the variance is a function of the mean using $V(y_i) = v(\mu_i) = \mu_i(1-\mu_i)$ where $\mu_i \in (0, 1)$ depends on the expected value of the response, and to ensure that the CDF integrates to 1, the normalizing constant is independent of the natural parameter. Being an EDM means it is easy to form estimating equations to estimate the β values.

Estimating the β values

Step 1

Rather than assuming an extremely restrictive distribution for the binary data (which restricts the maximum change to probability 0.5), we propose that a distribution such as the Burr type 10 distribution be chosen, as it allows for the inclusion of a disturbance term without strong assumptions of symmetry. This is advantageous in that, as shown in the equation, when $\phi = 1$, the distribution transforms itself to the logit. This means that our model is flexible in modeling both symmetrical and asymmetrical binomial data.

Step 2

Choose an initial $\phi^{(0)}$ for the skewness parameter to estimate the true ϕ . We use the estimate from the skewed logit regression from the binomial GLM, which includes the same fixed effect (model covariate) and different time intercepts. We also use the VCE to relax the assumption of independence inherent in the GLMs. The VCE is given by

$$(X'X)^{-1} \sum \mu'_i \mu_j (X'X)^{-1} \quad (20)$$

. Our data were collected over time and are therefore, not independent.

Step 3

Confirm that $kJ^{(-1)}k' < \epsilon$ where ϵ is a small constant, such as 0.001 . Repeat step 2 using these educated guesses. Once the conditions are satisfied, it means that convergence has been achieved and the estimated ϕ is the most robust estimate to be used in the GEE.

Step 4

From the Burr type 10 distribution,

$$F(k; \phi) = \frac{1}{(1 + \exp^{-k})^\phi}, k^{-1}(\mu_{ij}) = X_{ij}^T \quad (21)$$

will be the link function that relates the first moments to the covariates of interest.

Step 5

After obtaining the estimated value of ϕ , we use it to modify the skewness parameter for the binomial response, (usually assumed to be 1 in the GEEM package in the R software).

Use the estimated ϕ to update the estimated values of β .

Step 6

Run the GEEM model using morbidity as the response and BV, feeding group, HIV status, weight, and gender as the covariates. To obtain an adequate model, we systematically added the predictors in order of importance while updating the β values as shown by the equation. Once convergence was attained, the estimated β values and their standard errors were obtained considering significance at $p = 0.05$.

Appendix B: Proof for alternative calculation of scale parameter by altering the denominator under Exchangeable correlation

Scale parameter

Following [[Hardin, 2013](#)] the scale parameter also known as dispersion parameter is usually estimated as

$$\hat{\phi} = \frac{1}{(\sum_i^n n_i) - p} \sum_i^n \sum_i^{n_i} \hat{r}_{it}^2 \quad (22)$$

Where $\sum n_i$ is the total number of observation, \hat{r}_{it} is the i^{it} pearson residual and p is the total number of covariates used in the model.

To estimate α , [Hardin \[2013\]](#) calculates the denominator of an exchangeable correlation on page 98 as

$$\sum_i^{n_i} 0.5 \times n_i(n_i - 1) - p \quad (23)$$

We propose that the same can be implemented using combinations. We suggest the following

$$\sum_i^{n_i} n_i \text{Combn}, 2 - p \quad (24)$$

for $n_i > 1$ where n_i are intergers

Proof by induction

For balanced clusters, assume the data given by [Hardin and Hilbe page 66](#). The data has 2 subjects with 4 time periods. To calculate the denominator then it can be shown that

$$0.5 \times 4(4 - 1) + 0.5 \times 4(4 - 1) = 12 - p$$

Similarly using our method

$$4\text{Combn}, 2 = 6 \times 2 = 12 - p$$

Hence the proof.

We gain more efficiency in run time when we use our proposed approach than when we use Hardin and Hilbe approach as seen in Table 1

TABLE 1: *Run time in seconds in calculating the denominator of the scale parameter*

Method	Run time in seconds
Hardin and Hilbe	0.033
Proposed method 1	0.0319
Proposed method 2	0.0289

Appendix C: R codes

To create the Bivariate maps on mean and median cost with the household age

load the master shapefile>layers>add vector layers>select the master shapefile.

To split into different counties shapefiles.

click vector>datamanagement tools>split vector layer>select the required attribute(example are the counties)> on input leave blank>output directory, select the folder to put the split files.

finally after saving in a folder, if you want to bring in like 2 files ,Kitui number 30 and embu number 28, locate their files through the split numbers.

go to layers>import layers you want>this case the two counties

To prepare the files for use in R

Working with the QGIS to create shapefiles to use in R to produce spatial maps.

Download freely available shapefiles at <https://www.diva-gis.org/gdata> and save in a folder. This include .CPG, .DBF, .PRJ, .QPJ, .SHP, and .SHX. Open a new project in QGIS and import the shape file. This is done by clicking Layer>add layer> Add vector layer>vector. Then on the dialog box select the .SHP file into the project.

That command loads the shapefiles into the QGIS for use. But for us to get the correct format to enable us merge our prepared CSV with the counties data, we export as .CSV. Right click on the file>export>save feature as>on the format select, comma separated

value>dialogue opens, on file name input the extension to the folder>file name; leave the rest default then>ok.

Now you have the counties in form of .CSV. ensure your counties name are marching. Some like Keiyo-Marakwet are referred to as Elgeyo marakwet. Ensure your .CSV file with attributes such as distance to health facility, cost of pay, with the ID COUNTY. ; matches with the ID COUNTIES in the exported .CSV.

Go back to the QGIS, create new project>import the master shape files as above(originally downloaded shape files).

To add our .CSV file for merging purpose, click on Add layer>add delimited text layer>file name> select the minor .CSV with your data for bivariate analysis>add.

Right click on the master file>properties.....OR.....click on the master layer>layer properties.

At the bottom of that screen there is a + sign ; on the left something looks like a triangle shaded blue>click> dialogue box opens>join layer>select the minor layer>join field>select the unique id(county which are matching in both files)> on target field select unique id again. Ensure the master and the minor files have the same unique ids. Joined field select all the attributes from the minor that you would wish to plot the bivariates>click ok.

On top left of the dialogue screen you should see >join layer>value of the other to join>OK

To ensure you have the attributes correctly imported>right click on the master imported>open attribute table>ensure the needed attributes are there. If not repeat again. Export this file >select Esri shape file>select the file name to save, on the folder you are working with in R>on CRS select the EPSG:4326-WGS84>the rest are default then click ok. You should see all the files .CPG, .DBF, .PRJ, .QPJ, .SHP, and .SHX.with the name you gave them. The .SHP should have the largest size.

NB. Ensure that your .CSV are general in the numbers as you save. Else they will be read as factors in R.

To rename the headers in QGIS

Once you merge the QGIS shape files with the covariates of choice as calculated by the R program, then you may need to rename before exporting as shapefiles to use in R. Right click on the shapefile>properties>sourcefields[there are three icons on top, New field, delete field, toggle field, field calculator]> select toggle edit field>edit to the name of choice>you can save the shapefiles as you wish

R Code for Tables in section 4.2

```
rm(list = ls())##clean the memory
setwd("C:/Users/user/Desktop/restructureFRONTIERS")
library(geepack)# performing GEE regression
library(Greg)## displaying the odds ratio for GLM
library(haven)#for importing spss datasets
library(fmsb) #for odds ratio
library(dplyr)###for data wrangling and manipulation
library(lubridate)###for manipulating the dates
library(survminer)###for survival curves
library(survival)###for survival curves
library(KMsurv)###for survival curves
#####we create the bacterial vaginosis from the BV data.
any value greater than 7 then positive for BV
df1 <- read_sav("mbbv.sav") ##read the BV data
df1_1<-select(df1, IDNUM, BV) #subset df1 by selecting
the BV measure only. It's a pH measures between 0 and 14
df1_2<-df1_1[complete.cases(df1_1), ] #remove any NA or
Missing data from the dataset
df1_bv<-df1_2 %>% group_by(IDNUM)%>% summarise(BV=max(BV))
#--we want to work with the maximum BV values. assumption,
if you have tested positive for BV then you are exposed
```

```

df1_bv$IDNUM<-as.integer(df1_bv$IDNUM) #convert to interger
for joining with the morbidity dataset
head(df1_bv) ##the BV data with the pH values as intergers
#####
##### PART 2 #####
#####
####Next we read the infant morbidity dataset#####
####which we will merge with the BV dataset above#####
####then we analyse#####
#####
#####
mbinfant <- read_sav("mbinfant.sav")#read the infant morbidity data
mbneo <- read_sav("mbneo.sav") ##read the neonates data
mbneo$IWEIGHT<-mbneo$BIRTHWT/10 ##match weight in the two datasets
mbneo$HEIGHT<-mbneo$LENGTH #match height/length in the 2 datasets
#merge the data to obtain W00(week Zero/birth) in mbinant and
calculate the number of days
mbInfNeo<-merge(mbinfant, mbneo,
+ by =c("IDNUM","DAY","MONTH","YEAR","YEAR2K","VISIT","MB","MBNUM",
+ "PERIOD","TIME","IWEIGHT","HEIGHT"), all = TRUE, sort = TRUE)
mbneo2 <- read_sav("mbneo2.sav") ##load data with mortality information
##get the age at death/survival
mbneo2_dead<-mbneo2 %>% group_by(idnum,randgrp,deathage,dead,
+ pcrpos,deathage,monthage) %>% summarise()
mbneo2_dead_1<-mbneo2_dead %>% replace(is.na(.), -100)##replace NA with a
unique small number to easen filtering
##group by id and summarize
mbd2a<-mbneo2_dead_1 %>% group_by(idnum) %>% summarise(dead=max(dead),
+ hiv=max(pcrpos),rndgrp=max(randgrp),death_age=max(deathage),
+ month_age=max(monthage))# %>% max_age=pmax(deathage, monthsage)
##find survival age OR age at Death

```



```

mbd2a$max_age<-with(mbd2a, pmax(death_age, month_age)) #get age at
death/ survival upto the time
mbd2<-mbd2a %>% mutate(age2=max_age*30)##convert the age months into
days by multiplying by 30 days
#####
colnames(mbd2)[1]<-"IDNUM" #rename the first column from "idnum" to
"IDNUM" to match the other datasets
####join the selected data with deaths from mbne0 with the mbinfant
mbd2$IDNUM<-as.integer(mbd2$IDNUM)
dg1_1<-mbInfNeo %>% left_join(mbd2, by="IDNUM") ##use the function
left_join to merge the 2 datasets
##join the data with the BV data created above
dg1<-dg1_1 %>% left_join(df1_bv, by="IDNUM") #merge bv
#sorting with dates
dg1$Date <- with(dg1, dmy(sprintf('%02d%02d%04d',DAY,MONTH,YEAR2K)))#convert
the date into the R format of date
dg3<-dg1 %>% mutate(Date = ymd(Date))
dg4<-dg3[with(dg3,order(IDNUM,Date)),] ##ordering the dates in the data,
so that they are properly arranged.from birth upto the last visit
dg5<-dg4 %>% group_by(IDNUM) %>% mutate(diff2 = Date-first(Date))##find the
difference in days between the visits
##### PART 3 #####
####Next we Analyse data for the first 6months#####
#### data analysis for the 180 days(6months) #####
##the first 180 days
dg6<-dg5 %>% filter(diff2<180) ##filter to remain with data for 180days
dg6$bvyes<-ifelse(dg6$BV>=7,1,0)##if the pH is 7 or greater than 7
then they are positive for BV and 0 otherwise
dg6$bvyes[is.na(dg6$bvyes)] <- -100 ##recode missing to -100 for filtering
dg63<- dg6 %>% filter(bvyes>-1) ###filter all -100 which were missing
##our data for analysis is called dg63###

```

```
##### Table 1 #####
#### We use simple logistic regression #####
### Since there is no corelation to account for #
#merge the neonates data with the BV data from women to
calculate unadjusted odds ratios
mbneo <- read_sav("mbneo.sav") #read the neonates data
mbneo$IDNUM<-as.integer (mbneo$IDNUM) #convert NEONATES id to
interger for merging
df1_bv$IDNUM<-as.integer(df1_bv$IDNUM) #convert the mothers BV
data to interger
dh1<-mbneo %>% left_join(df1_bv, by="IDNUM") #join the data sets
dh1$BV[is.na(dh1$BV)]<--100 ##assign a large negative interger for
ease of filtering
dh2<- dh1 %>% filter(BV>-1) # filter all missing
dh2$bvys<-ifelse(dh2$BV>=7,1,0)# recode the pH to numeric
#####dealing with continuous variables in table 1 ####
dh2_0<-dh2 %>% filter(bvys==0)
dh2_1<-dh2 %>% filter(bvys==1)
mean(dh2_0$LENGTH, na.rm=TRUE)
mean(dh2_1$LENGTH, na.rm=TRUE)
stdev(dh2_0$LENGTH, na.rm=TRUE, unbiased=TRUE)
stdev(dh2_1$LENGTH, na.rm=TRUE, unbiased=TRUE)
range(dh2_0$LENGTH, na.rm=TRUE)
range(dh2_1$LENGTH, na.rm=TRUE)
length(dh2_0$MB)
## average BIRTHWT
mean(dh2_0$BIRTHWT, na.rm=TRUE)
mean(dh2_1$BIRTHWT, na.rm=TRUE)
range(dh2_0$BIRTHWT, na.rm=TRUE)
range(dh2_1$BIRTHWT, na.rm=TRUE)
stdev(dh2_0$BIRTHWT, na.rm=TRUE, unbiased=TRUE)
```

```
stdev(dh2_1$BIRTHWT, na.rm=TRUE, unbiased=TRUE)
#####no of days in hospital stay
dh2_0 <- dh2[ which(dh2$bvys=='0'),]
mean(dh2_0$HOSPSTAY, na.rm=TRUE)
stdev(dh2_0$HOSPSTAY, na.rm=TRUE, unbiased=TRUE)
median(dh2_0$HOSPSTAY, na.rm=TRUE)
summary(dh2_0$HOSPSTAY, na.rm=TRUE)
range(dh2_0$HOSPSTAY, na.rm=TRUE)
sum(dh2_0$HOSPSTAY>0, na.rm = TRUE)
sum(dh2_1$HOSPSTAY>0, na.rm = TRUE)
dh2_1 <- dh2[ which(dh2$bvys=='1'),]
mean(dh2_1$HOSPSTAY, na.rm=TRUE)
stdev(dh2_1$HOSPSTAY, na.rm=TRUE, unbiased=TRUE)
range(dh2_1$HOSPSTAY, na.rm=TRUE)
##more than 24 hours in hospital
sum(dh2_0$HOSP24HR, na.rm=TRUE)
sum(dh2_1$HOSP24HR, na.rm=TRUE)
###nsephis
sum(dh2_0$NSEPSIS, na.rm=TRUE)
sum(dh2_1$NSEPSIS, na.rm=TRUE)
##MATCOND
sum(dh2_0$MATCOND, na.rm=TRUE)
sum(dh2_1$MATCOND, na.rm=TRUE)
##NRASH
sum(dh2_0$NRASH, na.rm=TRUE)
sum(dh2_1$NRASH, na.rm=TRUE)
#NLYMPHAD
sum(dh2_0$NLYMPHAD, na.rm=TRUE)
sum(dh2_1$NLYMPHAD, na.rm=TRUE)
##DISTRESS
sum(dh2_0$DISTRESS, na.rm=TRUE)
```

```
sum(dh2_1$DISTRESS, na.rm=TRUE)
## head circumference
mean(dh2_0$NHCIRC, na.rm=TRUE)
mean(dh2_1$NHCIRC, na.rm=TRUE)
range(dh2_0$NHCIRC, na.rm=TRUE)
range(dh2_1$NHCIRC, na.rm=TRUE)
stdev(dh2_0$NHCIRC, na.rm=TRUE, unbiased=TRUE)
stdev(dh2_1$NHCIRC, na.rm=TRUE, unbiased=TRUE)
##apgar score
mean(dh2_0$APGAR, na.rm=TRUE)
mean(dh2_1$APGAR, na.rm=TRUE)
range(dh2_0$APGAR, na.rm=TRUE)
range(dh2_1$APGAR, na.rm=TRUE)
stdev(dh2_0$APGAR, na.rm=TRUE, unbiased=TRUE)
stdev(dh2_1$APGAR, na.rm=TRUE, unbiased=TRUE)
##dubowitz score
mean(dh2_0$DUBOWITZ, na.rm=TRUE)
mean(dh2_1$DUBOWITZ, na.rm=TRUE)
range(dh2_0$DUBOWITZ, na.rm=TRUE)
range(dh2_1$DUBOWITZ, na.rm=TRUE)
stdev(dh2_0$DUBOWITZ, na.rm=TRUE, unbiased=TRUE)
stdev(dh2_1$DUBOWITZ, na.rm=TRUE, unbiased=TRUE)
##maturity
mean(dh2_0$MATURITY, na.rm=TRUE)
mean(dh2_1$MATURITY, na.rm=TRUE)
range(dh2_0$MATURITY, na.rm=TRUE)
range(dh2_1$MATURITY, na.rm=TRUE)
stdev(dh2_0$MATURITY, na.rm=TRUE, unbiased=TRUE)
stdev(dh2_1$MATURITY, na.rm=TRUE, unbiased=TRUE)
##jaundice
sum(dh2_0$NJAUNDH, na.rm=TRUE)
```

```

sum(dh2_1$NJAUNDH, na.rm=TRUE)
##conjuviticus
sum(dh2_0$NCONJUNC, na.rm=TRUE)
sum(dh2_1$NCONJUNC, na.rm=TRUE)
#####glm model for neonates
##run a simple logistic model to extract unadjusted odds ratio
mod4<-glm(bvys~DISTRESS+NSEPSIS+NRASH+NLYMPHAD+JAUNDICE+NCONJUNC,
+ data =dh2,family = "binomial")
printCrudeAndAdjustedModel(mod4)[-1,] #print odds ratio less intercept
#### Table 2 for the 180 days #####
#### We use GEE with independencecorrelation #####
### Since there is corelation within the infant to account for ##
#create a subset of the main data dg63 with variables of analysis
dg63_2<-dg63%>% select(IDNUM,bvyes,IPNEUM,EARINFEC,STOOLBLD,
      ILYMPHAD,IDIARMON,ENCEPHAL,ISEPSIS,ICONJUNC,DEHYDRAT,
      GEDIARRH,GEVOMIT,WHEEZING,IHEPATOM,COLD,OTITIS,IHAIRYLP,
      STOOLBLD,IHOSPIT,CLINIC,IFEVER,ICOUGH,IDIARHEA,ORSLW,ITHRUSH,
      VOMIT,FEEDDIFF,HEATRASH,FUNGRASH,ECDERMAT,SCABIES,IORALULC)
dataModel1<-na.omit(dg63_2)##remove all missing data to work well with GEE
modGEE<-geeglm(bvyes~IPNEUM+EARINFEC+STOOLBLD+
      ILYMPHAD+ENCEPHAL+ISEPSIS+ICONJUNC+DEHYDRAT+
      GEDIARRH+GEVOMIT+WHEEZING+IHEPATOM+COLD+OTITIS+
      STOOLBLD+IHOSPIT+CLINIC+IFEVER+ICOUGH+IDIARHEA+ORSLW+ITHRUSH+
      VOMIT+FEEDDIFF+HEATRASH+FUNGRASH+ECDERMAT+SCABIES+IORALULC,
      family=binomial(link="logit"),data =dataModel1,id = IDNUM,
      corstr = "independence")
summary(modGEE)
coefi<-summary(modGEE)$coefficients[, 1]##extract model coefficients
se.err<-summary(modGEE)$coefficients[, 2] ##extract the standard errors
to calculate the confidence interval
oddsRatio<-exp(coefi) ##calculate the odds ratio by exponentiating

```

```

the coefficients
lowerCI<-exp(coefi-1.96*se.err) ##2.5% lower confidence interval
upperCI<-exp(coefi+1.96*se.err) ##97.5% upper bound
oddsGEE<-cbind(oddsRatio,lowerCI,upperCI) #bind all the data together
oddsGEE  ##the output
###using GLM to compare the results
###but since our method is about GEE, we only report results from the GEE output
modGLM<-glm(bvyes~IPNEUM+EARINFEC+STOOLBLD+
            ILYMPHAD+ENCEPHAL+ISEPSIS+ICONJUNC+DEHYDRAT+
            GEDIARRH+GEVOMIT+WHEEZING+IHEPATOM+COLD+OTITIS+
            STOOLBLD+IHOSPIT+CLINIC+IFEVER+ICOUGH+IDIARHEA+ORSLW+ITHRUSH+
            VOMIT+FEEDDIFF+HEATRASH+FUNGRASH+ECDERMAT+SCABIES+IORALULC,
            data =dataModel1,family = "binomial")
summary(modGLM)
exp(cbind("Odds ratio" = coef(modGLM), confint.default(modGLM,
+ level = 0.95)))##calculate odds ratio from model adjusted
printCrudeAndAdjustedModel(modGLM)[-1,] ##both unadjusted and adjusted odds ratio

```

Additionally R Code for Models in section 4.3 for comparing Logit and Skewed Logit under GEE

```

#finding the skewness parameter using stata scobit function
#scobit fev time weight brestfed male hiv_infected bv##
time ,vce(cluster idnum) nrtol(1e-3)
#####start here
library(geeM)
#library(geepack)
modelData <- read.csv("C:/Users/user/Desktop/trash2/longd2.txt")
k <- 1
linkfun <- function(p){log((p^(1/k))/(1 - p^(1/k)))}
variance <- function(p){p * (1-p)}

```

```

linkinv <- function(eta){(exp(eta)/(1 + exp(eta)))^k}
mu.eta<-function(eta){k*(exp(eta))^(k-1)/(1+exp(eta))^(k+1)}
FunList <- list(linkfun, variance, linkinv, mu.eta)
model_GEE<-geem(fev~bv*time+male+BRESTFED+HIV_INFECTED+weight,
data =modelData,id=IDNUM,family = FunList,corstr="ar1")
summary(model_GEE)

#model SGEE after substarcting the skewness parameter from 1.
# 1 is normal
k <- 0.9224
linkfun <- function(p){log((p^(1/k))/(1 - p^(1/k)))}
variance <- function(p){p * (1-p)}
linkinv <- function(eta){(exp(eta)/(1+exp(eta)))^k}
mu.eta <- function(eta){k*(exp(eta))^(k-1)/(1+exp(eta))^(k+1)}
FunList <- list(linkfun, variance, linkinv, mu.eta)
model_SGEE<-geem(fev~bv*time+male+BRESTFED+HIV_INFECTED+weight,
data =modelData,id=IDNUM,family = FunList,corstr="ar1")
summary(model_SGEE)

```

R code for section 4.4 on Data

```

rm(list = ls())#clear any data in memory
library(haven)
library(dplyr)
library(haven)
library(stats)
library(statmod)
library(tweedie)
library(labelled)
library(foreign)
library(purrr)
library(ggplot2)

```

```
library(e1071)
library(sf)
library(biscale)
library(ggplot2)
library(cowplot)
library(sp)
library(rgdal)
library(tmap)
library(dplyr)
library(leaflet)
##analysis starts here
df1 <- read_sav("C:/redd/paper3/tweedie_kheus_data/
Rcodes2013/kheus2018/khheus_c1.sav")
attach(df1)
d2<-df1 %>%
  select(county,resid,clid,hhid,s_num,wealth_index1,q3,q4,q5,q6b,q8,
q10,q11,q14,q15a,q15b,q15c,q15d,q15e,q15f,q15g,q15h,q15i,q36tot)%>%
  rename(county=county,res=resid,clNum=clid,hh=hhid,vis_num=s_num,
wealthIndx=wealth_index1,rel_Head=q3,sex=q4,religion=q5,age=q6b,
high_educ=q8,mar_status=q10,empl_stats=q11,smoker=q14,hypert=q15a,
cardiac=q15b,diabetes=q15c,asthma=q15d,TB=q15e,other_respiratory=q15f,
HIV=q15g,cancer=q15h,mental=q15i,total_cost=q36tot)
data<-d2 %>% filter(s_num==1)##filter to remain with 1 visit only
##household head information
f1<-data %>% group_by(clNum,hh)%>% filter(rel_Head==min(rel_Head))%>%
  filter(age==max(age))
d4<-f1 %>% select(clNum,hh,sex,religion,age,high_educ,
  mar_status,empl_stats)
head(d4)%>% as.data.frame()
#Household information
d5<-data %>% group_by(clNum,hh)%>%
```



```

summarise(res=max(res),wlth_index=max(wealthIndx),
smoker=min(smoker),hypert=min(hypert),cardiac=min(cardiac),
diabetes=min(diabetes),asthma=min(asthma),TB=min(TB),
any_respiratiry=min(other_respiratory), HIV=min(HIV),
cancer=min(cancer), mental=min(mental),
totSpend=sum(total_cost, na.rm = TRUE))
d6<-d4 %>% left_join(d5,by=c("c1Num","hh")) %>% as.data.frame()
d7 =d6%>% filter(age>17)%>% filter(mar_status<5 )
d8<-dplyr::select(d7, -c1Num,-hh,-religion)
##create variables
d8$high_educ1<-ifelse(d8$high_educ==8|d8$high_educ==9,0,
  ifelse(d8$high_educ==1|d8$high_educ==2|d8$high_educ==7,1,
  ifelse(d8$high_educ==3|d8$high_educ==6,2,3)))
d8$empl_stats1<-ifelse(d8$empl_stats==1|d8$empl_stats==2,1,0)
d9<-d8 %>% filter(smoker<3)%>% filter(hypert<3)%>%
filter(cardiac<3)%>% filter(diabetes<3)%>% filter(asthma<3)%>%
filter(TB<3)%>% filter(any_respiratiry<3)%>% filter(HIV<3)%>%
filter(cancer<3)%>% filter(mental<3)
d10<-dplyr::select(d9, -high_educ,-empl_stats)
power2=tweedie.profile(totSpend~age+factor(wlth_index)+
factor(mar_status)+factor(high_educ1),
  p.vec=seq(1.5,1.8,length=10),
  do.ci=TRUE, method="interpolation", data = d10)
p1=power2$p.max
power2$ci
glmmodel1<-glm(totSpend~age+factor(mar_status)+factor(wlth_index)+factor(high_educ1
  family=tweedie(var.power=p1, link.power=0),x=TRUE, data=d10)
fits1<-glmmodel1$fitted.values
beta=glmmodel1$coefficients
phi = power2$phi.max
n=length(d10$totSpend)

```

```

r1=glmmodel1$rank
quasi<-sum((totSpend*fits1^(1-p1)/(1-p1))-((fits1^(2-p1))/(2-p1)))
qicu<-(-2*quasi)+(2*r1)
qicu
# Finding a suitable link function
glmmodel<-glm(totSpend~age+factor(wlth_index)+factor(mar_status)+
factor(high_educ1),
family=tweedie(var.power=p1, link.power=0),x=TRUE, data=d10)
glmmodel.other<-glm(totSpend~age+factor(wlth_index)+
factor(mar_status)+factor(high_educ1),
family=tweedie(var.power=p1),x=TRUE, data = d10) # Canonical
#Deviances
glmmodel$deviance
glmmodel.other$deviance
#Df Residuals
glmmodel$df.residual
glmmodel.other$df.residual
dk1<-data %>% group_by(age) %>%
summarise(tt_cost=mean(total_cost,na.rm = TRUE))
dk2<-d10 %>% group_by(age) %>%
summarise(tt_cost=mean(totSpend,na.rm = TRUE),n=n())%>%as.data.frame()
dk3<-dk2 %>% filter(age<97)
ggplot(dk2) +
  geom_line( mapping = aes(x = age, y = tt_cost))
dk3$x<-dk3$age
dk3$y<-dk3$tt_cost
plot(dk3$x, dk3$y, type = "l", lty = 1,
xlab="Age in Years of Household Head",
ylab = "Mean Cost for Outpatient Care per Household")
lines(dk3$x, dk3$y, type = "l", lty = 1)
sd(d10$totSpend)

```

```
mean(d10$totSpend)
median(d10$totSpend)
skewness(d10$totSpend)
max(d10$totSpend)
min(d10$totSpend)
d11<-d10 %>% filter(totSpend>0)
sd(d11$totSpend)
mean(d11$totSpend)
median(d11$totSpend)
skewness(d11$totSpend)
max(d11$totSpend)
min(d11$totSpend)
##bivariate spatial maps
head(data)
dh1<-data %>% select(county,age,total_cost)
dh2<-dh1 %>% group_by(county) %>%
summarise(age1=mean(age, na.rm = TRUE),
tot_mean=mean(total_cost,na.rm = TRUE),
age2=median(age,na.rm = TRUE),tot_med=median(total_cost,na.rm = TRUE))
dh3<-dh1 %>% filter(age>17) %>% filter(age<97)
dh21<-dh3 %>% group_by(county) %>%
summarise(age1=mean(age, na.rm = TRUE),
tot_mean=mean(total_cost,na.rm = TRUE),
age2=median(age,na.rm = TRUE),tot_med=median(total_cost,na.rm = TRUE))

##for producing maps
fg<-"C:.....county.shp"
fg
nc <- st_read(fg)
head(nc)
data <- bi_class(data, x = Mean_age,
```

```
y = mean_cost, dim = 3)
head(data)
ggplot() +
  geom_sf(data = data, aes(fill = bi_class),
color = "red", size = 0.1, show.legend = TRUE) +
  bi_scale_fill(pal = "GrPink", dim = 3) +
  bi_theme()
legend <- bi_legend(pal = "GrPink", dim = 3,
xlab = "Mean age of Household head",
ylab = "Mean outpatient cost(KES)", size = 8)
plot <- ggdraw() +
  draw_plot(map, 0, 0, 1, 1) +
  draw_plot(legend, 0.1, 0.1, 0.2, 0.2)
plot
```

R code for section 4.5: Part 1 on Data cleaning

```
# Initial setting of data
# Load libraries that are needed to perform calculations
rm(list = ls())
library(dplyr)
library(haven)
library(stats)
library(statmod)
library(tweedie)
library(labelled)
library(foreign)
#----set working directory
setwd('C:/redd/paper3/tweedie_kheus_data/Rcodes2013')
#-----set sink() directory
sink('sink/inpatients_out.txt')
#---read the data. Note all the output regarding the data are
put in the sink() file
#dh<- read_sav("kheus2013_data_questionnaire/In-patients_Data.sav")
dj<-read.spss("kheus2013_data_questionnaire/In-patients_Data.sav",
+ to.data.frame=TRUE)
View(dj) ##View the data
head(dj) # first few variables but in the .txt file
##incase you want to deactivate the sink, uncoment on the sink(below)
#sink() ##deactivating the sink
#####-----End of thread
#some analysis using dplyr
##sumarise the number of times county
a1<-dj %>% group_by(CountyCode) %>% summarise(number = n())
##sumarise the number of times of hospital visits
a2<-dj %>% group_by(Q50) %>% summarise(number = n())
#amount paid for 6 admissions
```

```

sixAdmPay<-dj %>% filter(Q50==6)
#how long was name admitted
n_admit<-dj %>%group_by(Q51)%>% summarise(n=n())
####type of facility and ownership(the one visited)---
#-OWNERSHIP OF THE FACILITIES GROUPEd
l_hosp<-dj %>% group_by(Q53) %>% summarise(n=n())
dj$fac_owned_govt<-ifelse(dj$Q53=="Govt. Hospitals"
+ |dj$Q53=="Govt. Health Centre",1,0)
dj$fac_owned_private<-ifelse(dj$Q53=="Private hospitals"
+ |dj$Q53=="Private Health Centre"|dj$Q53=="Nursing/Maternity Homes",1,0)
dj$fac_owned_mission<-ifelse(dj$Q53=="Mission Hospital"
+ |dj$Q53=="Mission health centre",1,0)
#####-----end of clasifying
##counties where they visited traditional healer
trad_healer<-dj %>% filter(Q53=="Traditional healer")
##is this the nearest inpatient facility?
near_facilit<-dj %>%group_by(Q54)%>% summarise(n=n())
####type of facility and ownership(the one close to home)
l_hosp2<-dj %>% group_by(Q55) %>% summarise(n=n())
#reason for passing the facility
####why pass facility(reason 1) arrange in descending order
reas_passA<-dj %>% group_by(Q56A) %>% summarise(n=n())%>% arrange(desc(n))
####why pass facility(reason 2) arrange in descending order
reas_passB<-dj %>% group_by(Q56B) %>% summarise(n=n())%>% arrange(desc(n))
####why pass facility(reason 3) arrange in descending order
reas_passC<-dj %>% group_by(Q56C) %>% summarise(n=n())%>% arrange(desc(n))
#reason for choosing the facility---some analysis to selects the 5main
#reasons for CHOOSING THE HEALTH FACILITY
####why pass facility(reason 1) arrange in descending order
reas_chooseA<-dj %>% group_by(Q57A) %>% summarise(n=n())%>% arrange(desc(n))
####why pass facility(reason 2) arrange in descending order

```

```

reas_chooseB<-dj %>% group_by(Q57B) %>% summarise(n=n())%>% arrange(desc(n))
####why pass facility(reason 3) arrange in descending order
reas_chooseC<-dj %>% group_by(Q57C) %>% summarise(n=n())%>% arrange(desc(n))
reas1<-dj %>% select(Q57A)%>% rename(reas=Q57A)
reas2<-dj %>% select(Q57B)%>% rename(reas=Q57B)
reas3<-dj %>% select(Q57C)%>% rename(reas=Q57C)
total <- rbind(reas1,reas2,reas3)
re<-total %>% group_by(reas) %>% summarise(n=n())%>% arrange(desc(n))
#create a binary close to home as reason for choosing the facility
dj$reason_closehome<-ifelse(dj$Q57A=="Close to home"
+ |dj$Q57B=="Close to home"|dj$Q57C=="Close to home",1,0)
dj$reason_closehome[is.na(dj$reason_closehome)]<-0
dj$reason_staffQualified<-ifelse(dj$Q57A=="Staff are qualified"
+ |dj$Q57B=="Staff are qualified"|dj$Q57C=="Staff are qualified",1,0)
dj$reason_staffQualified[is.na(dj$reason_staffQualified)]<-0
dj$reason_medicineavailable<-ifelse(dj$Q57A=="Medicine available"
+ |dj$Q57B=="Medicine available"|dj$Q57C=="Medicine available",1,0)
dj$reason_medicineavailable[is.na(dj$reason_medicineavailable)]<-0
dj$reason_Less_costly<-ifelse(dj$Q57A=="Less costly" |dj$Q57B=="Less costly"
+ |dj$Q57C=="Less costly",1,0)
dj$reason_Less_costly[is.na(dj$reason_Less_costly)]<-0
dj$reason_Goodstaffattitude<-ifelse(dj$Q57A=="Good staff attitude"
+ |dj$Q57B=="Good staff attitude"|dj$Q57C=="Good staff attitude",1,0)
dj$reason_Goodstaffattitude[is.na(dj$reason_Goodstaffattitude)]<-0
dj$reason_Wasreferred<-ifelse(dj$Q57A=="Was referred"
+ |dj$Q57B=="Was referred"|dj$Q57C=="Was referred",1,0)
dj$reason_Wasreferred[is.na(dj$reason_Wasreferred)]<-0
dj$reason_Lesswaitingtime<-ifelse(dj$Q57A=="Less waiting time"
+ |dj$Q57B=="Less waiting time"|dj$Q57C=="Less waiting time",1,0)
dj$reason_Lesswaitingtime[is.na(dj$reason_Lesswaitingtime)]<-0
dj$reason_Moreprivacy<-ifelse(dj$Q57A=="More privacy"

```

```

+ |dj$Q57B=="More privacy"|dj$Q57C=="More privacy",1,0)
dj$reason_Moreprivacy[is.na(dj$reason_Moreprivacy)]<-0
####-----END of reasons FOR CHOOSING FACILITY
####reason for seeking inpatient services at the facility---REASON SICKNESS
reas_admitA<-dj %>% group_by(Q58A) %>% summarise(n=n())%>% arrange(desc(n))
reas_admitB<-dj %>% group_by(Q58B) %>% summarise(n=n())%>% arrange(desc(n))
reas_admitC<-dj %>% group_by(Q58C) %>% summarise(n=n())%>% arrange(desc(n))
sicknes1<-dj %>% select(Q58A)%>% rename(sick=Q58A)
sicknes2<-dj %>% select(Q58B)%>% rename(sick=Q58B)
sicknes3<-dj %>% select(Q58C)%>% rename(sick=Q58C)
sickt <- rbind(sicknes1,sicknes2,sicknes3)
re2<-sickt %>% group_by(sick) %>% summarise(n=n())%>% arrange(desc(n))
#create a binary for sickness
dj$Malaria_fever<-ifelse(dj$Q58A=="Malaria/fever" |dj$Q58B=="Malaria/fever"
+ |dj$Q58C=="Malaria/fever",1,0)
dj$Malaria_fever[is.na(dj$Malaria_fever)]<-0
dj$respiratory_pneumonia<-ifelse(
dj$Q58A=="Diseases of Respiratory including pneumonia"
+ |dj$Q58B=="Diseases of Respiratory including pneumonia"
+ |dj$Q58C=="Diseases of Respiratory including pneumonia",1,0)
dj$respiratory_pneumonia[is.na(dj$respiratory_pneumonia)]<-0
dj$normal_delivery<-ifelse(dj$Q58A=="normal delivery"
+ |dj$Q58B=="normal delivery"|dj$Q58C=="normal delivery",1,0)
dj$normal_delivery[is.na(dj$normal_delivery)]<-0
dj$Accidents_and_injuries<-ifelse(dj$Q58A=="Accidents and injuries"
+ |dj$Q58B=="Accidents and injuries"|dj$Q58C=="Accidents and injuries",1,0)
dj$Accidents_and_injuries[is.na(dj$Accidents_and_injuries)]<-0
dj$Hypertension<-ifelse(dj$Q58A=="Hypertension" |dj$Q58B=="Hypertension"
+ |dj$Q58C=="Hypertension",1,0)
dj$Hypertension[is.na(dj$Hypertension)]<-0
dj$Diarrhoea<-ifelse(dj$Q58A=="Diarrhoea" |dj$Q58B=="Diarrhoea"

```



```

+ |dj$Q58C=="Diarrhoea",1,0)
dj$Diarrhoea[is.na(dj$Diarrhoea)]<-0
dj$caesarean<-ifelse(dj$Q58A=="caesarean" |dj$Q58B=="caesarean"
+ |dj$Q58C=="caesarean",1,0)
dj$caesarean[is.na(dj$caesarean)]<-0
dj$Diabetes<-ifelse(dj$Q58A=="Diabetes" |dj$Q58B=="Diabetes"
+ |dj$Q58C=="Diabetes",1,0)
dj$Diabetes[is.na(dj$Diabetes)]<-0
mode.pay<-dj %>% group_by(Q61F) %>% summarise(n=n())
##change the factors into characters
dj$Q61A<-as.character(dj$Q61A)
dj$Q61B<-as.character(dj$Q61B)
dj$Q61C<-as.character(dj$Q61C)
dj$Q61D<-as.character(dj$Q61D)
dj$Q61E<-as.character(dj$Q61E)
dj$Q61F<-as.character(dj$Q61F)
dj$pay_cash<-ifelse(dj$Q61A=="Cash"|dj$Q61B=="Cash"
+ |dj$Q61C=="Cash" |dj$Q61D=="Cash" |dj$Q61E=="Cash"|dj$Q61F=="Cash",1,0)
dj$pay_cash[is.na(dj$pay_cash)]<-0
dj$NHIF_and_other_methods<-
ifelse(dj$Q61A=="Community health insurance scheme"
+ |dj$Q61B=="Community health insurance scheme"|
      dj$Q61C=="Community health insurance scheme" |
+ dj$Q61D=="Community health insurance scheme"|
      dj$Q61E=="Community health insurance scheme"|
+ dj$Q61F=="Community health insurance scheme"|
dj$Q61A=="Given opportunity to pay later (credit)"|
dj$Q61B=="Given opportunity to pay later (credit)"|
dj$Q61C=="Given opportunity to pay later (credit)" |
dj$Q61D=="Given opportunity to pay later (credit)"|
dj$Q61E=="Given opportunity to pay later (credit)"|

```

```

dj$Q61F=="Given opportunity to pay later (credit)"|
dj$Q61A=="Waived/exempted"|dj$Q61B=="Waived/exempted"|
dj$Q61C=="Waived/exempted" |dj$Q61D=="Waived/exempted"|
dj$Q61E=="Waived/exempted"|dj$Q61F=="Waived/exempted" |
dj$Q61A==" Paid in kind"|dj$Q61B==" Paid in kind"|
dj$Q61C==" Paid in kind" |dj$Q61D==" Paid in kind"|
dj$Q61E==" Paid in kind"|dj$Q61F==" Paid in kind" |
dj$Q61A==" National Hospital Insurance Fund (NHIF)"|
dj$Q61B==" National Hospital Insurance Fund (NHIF)"|
dj$Q61C==" National Hospital Insurance Fund (NHIF)" |
dj$Q61D==" National Hospital Insurance Fund (NHIF)"|
dj$Q61E==" National Hospital Insurance Fund (NHIF)"|
dj$Q61F==" National Hospital Insurance Fund (NHIF)"|
dj$Q61A=="Private health insurance"|
dj$Q61B=="Private health insurance"|
dj$Q61C=="Private health insurance" |
dj$Q61D=="Private health insurance"|
  dj$Q61E=="Private health insurance"|
dj$Q61F=="Private health insurance",1,0)
dj$NHIF_and_other_methods[is.na(dj$NHIF_and_other_methods)]<-0
dpk1<-dj %>% select(Q61A,Q61B,Q61C,Q61D,Q61E,Q61F,pay_cash,NHIF_and_other_methods)
pay<-dj %>% group_by(Q59) %>% summarise(n=n())
## total pay for the service; sum all the breakdowns, q60_1 to q60_7;
including drugs, admission...
dj<-dj %>% mutate(sumTotPay = rowSums(cbind(Q60_1, Q60_2, Q60_3,
+ Q60_4, Q60_5, Q60_6, Q60_7), na.rm = T))
##replace the total sum above, by creating a new variable where if
total q60_8 ismissing, replace with above value
dj$sumTot<-ifelse(is.na(dj$Q60_8),dj$sumTotPay,dj$Q60_8)
#provider drugs and clinical service
drug_clin<-dj %>% group_by(Q64) %>% summarise(n=n())%>% arrange(desc(n))

```

```

#overall satisfaction
overall_satif<-dj %>% group_by(Q65) %>% summarise(n=n())%>% arrange(desc(n))
####-----dealing with time and distance
##time to admission as a tweedie-----starts here
dj$timeToAdmission_Tweedie<-ifelse(is.na(dj$Q67ah),0,dj$Q67ah)
##time to admission as a tweedie-----Ends here
##time to admission as continuous-----starts here
dj$timeAdmi_hour_to_minutes<-dj$Q67ah*60
##add the minutes of Q67am and the new coputed minutes from Q67ah
dj<-dj %>% mutate(timeMinutesAddmision =
+ rowSums(cbind(timeAdmi_hour_to_minutes, Q67am), na.rm = T))
###replace the NA in the computed variable with zero
dj$timeMinutesAddmision[is.na(dj$timeMinutesAddmision)]<-0
##time to admission as continuous-----Ends here
##Time to arrive at the facility as a tweedie-----starts here
dj$timeToArrive<-ifelse(is.na(dj$Q67bh),0,dj$Q67bh)
##Time to arrive at the facility as a tweedie-----Ends here
##Time to arrive at the facility as continuous-----starts here
dj$timeArrive_facility_hour_to_minutes<-dj$Q67bh*60
##add the minutes of Q67am and the new coputed minutes from Q67ah
dj<-dj %>% mutate(timeMinutesArriveFacility =
+ rowSums(cbind(timeArrive_facility_hour_to_minutes, Q67bm), na.rm = T))
###replace the NA in the computed variable with zero
dj$timeMinutesArriveFacility[is.na(dj$timeMinutesArriveFacility)]<-0
##Time to arrive at the facility as continuous-----Ends here
##dealing with DISTANCE
dj$newDistBelowOneKm<-ifelse(dj$Q68<=1,0,dj$Q68)
dj$newDistBelowOneKm[is.na(dj$newDistBelowOneKm)]<-0
##### end of exploatory data analysis-----end
###Q69 those who didnt pay any fare,NA with zero replace
dj$Q69[is.na(dj$Q69)]<-0

```

```

###create a dataset with the covariates variables of choice
data<-dj %>% select(newDistBelowOneKm,Q69,timeMinutesAddmision,w_index,hhsize,
                    gender,age,Q61A,Q65,Q57A,Q58A,Q56A,Q55,Q53,Q54,rurb,
                    fac_owned_govt,fac_owned_private,fac_owned_mission,
                    reason_closehome,reason_staffQualified,reason_medicineavailable,
                    reason_Less_costly,reason_Goodstaffattitude,reason_Wasreferred,
                    reason_Lesswaitingtime,reason_Moreprivacy,Malaria_fever,
                    respiratory_pneumonia,normal_delivery,Accidents_and_injuries,
                    Hypertension,Diarrhoea,caesarean,Diabetes,pay_cash,
                    NHIF_and_other_methods,CountyCode)%>%
  rename(costPayFareToFacility=Q69,satisfiedWithService=Q65,
         mainReasonChoseFacility=Q57A,reasonAdmision=Q58A,
         whoOwnsTheFacilityNearHome=Q55,reasonPassingNearFac=Q56A,
         whoOwnsFacilityYouVisited=Q53,isThisNearestFac=Q54,
         modeOfPayment=Q61A)
colSums(is.na(dk))
data<-data%>% filter(newDistBelowOneKm<999)
data2<-data %>% select(newDistBelowOneKm,CountyCode,costPayFareToFacility,
                    timeMinutesAddmision,w_index,hhsize,gender,age,modeOfPayment,rurb,
                    fac_owned_govt,fac_owned_private,fac_owned_mission,reason_closehome,
                    reason_staffQualified,reason_medicineavailable,reason_Less_costly,
                    reason_Goodstaffattitude,reason_Wasreferred,reason_Lesswaitingtime,
                    reason_Moreprivacy,Malaria_fever,respiratory_pneumonia,normal_delivery,
                    Accidents_and_injuries,Hypertension,Diarrhoea,caesarean, Diabetes,
                    pay_cash,NHIF_and_other_methods)
colSums(is.na(data2)) #summing all missing data
data3<-na.omit(data2) #remove all rows with missing data
length(data3$newDistBelowOneKm)##3167 observations
##-----main data is data3

```

R code for proof by induction and table 1

```

##R code for the proof using Hardin and Hilbe Approach
library(dplyr)
fun_M1<- function() {
  id<-c(1,1,1,1,2,2,2,2)
  t<-c(1,2,3,4,1,2,3,4)
  y<-c(4,5,6,7,5,6,7,8)
  x<-c(0,1,0,1,0,1,0,1)
  fd<-data.frame(id,t,y,x)
  m1<-glm(y~x,x=TRUE, data = fd)
  y.res=m1$residuals
  n=length(fd$id)
  new.phi2<-sum(y.res^2)/(n-r)
  res_total_n_for_alph<-fd %>%
    group_by(id) %>%
    summarise(resd = sum(combn(y.res, 2, FUN = prod)),n = n())%>%
    mutate(sampl_n = 0.5*n*(n-1))
  sum_res_alpha<-sum(res_total_n_for_alph$resd)
  sum_n_alpha<-sum(res_total_n_for_alph$sampl_n)
  alpha2<-(1/new.phi2)*(1/(sum_n_alpha-r))*sum_res_alpha
  alpha2 }
  start_time <- Sys.time()
  fun_M1()
  end_time <- Sys.time()
  end_time - start_time
##R code for the proof using our proposed method 1
fun_M1<- function() {
  fd<-data.frame(id,t,y,x)
  m1<-glm(y~x,x=TRUE, data = fd)
  y.res=m1$residuals
  n=length(fd$id)

```

```

new.phi2<-sum(y.res^2)/(n-r)
res_total_n_for_alph<-fd %>%
  group_by(id) %>%
  summarise(resd = sum(combn(y.res, 2, FUN = prod)),n = n())%>%
  mutate(sampl_n = choose(n, 2))
sum_res_alpha<-sum(res_total_n_for_alph$resd)
sum_n_alpha<-sum(res_total_n_for_alph$sampl_n)
alpha2<-(1/new.phi2)*(1/(sum_n_alpha-r))*sum_res_alpha
alpha2 }

start_time <- Sys.time()
fun_M1()
end_time <- Sys.time()
end_time - start_time

##R code for the proof using our proposed method 2
library(dplyr)
fun_M1<- function() {
fd<-data.frame(id,t,y,x)
m1<-glm(y~x,x=TRUE, data = fd)
y.res=m1$residuals
n=length(fd$id)
new.phi2<-sum(y.res^2)/(n-r)
res_total_n_for_alph<-fd %>%
  group_by(id) %>%
  summarise(resd = sum(combn(y.res, 2, FUN = prod)),n = n())%>%
  mutate(sampl_n = n*(n-1)/2)
sum_res_alpha<-sum(res_total_n_for_alph$resd)
sum_n_alpha<-sum(res_total_n_for_alph$sampl_n)
alpha2<-(1/new.phi2)*(1/(sum_n_alpha-r))*sum_res_alpha
alpha2 }

start_time <- Sys.time()
fun_M1()

```

```
end_time <- Sys.time()
end_time - start_time
```

R code for Raw residuals versus the observation numbers plot

4.5

```
par(mfrow = c(2, 2))
attach(d2)
#-----model 7
var.power=tweedie.profile(dist~factor(emp_cat)+factor(paid_cat)+
                           factor(hhsize_group),
                           p.vec=seq(1.4,1.75,length=10),
                           do.ci=TRUE, method="saddlepoint", data = d2)
var.p=var.power$p.max
modGlm<-glm(dist~factor(emp_cat)+factor(paid_cat)+
             factor(hhsize_group),
             family=tweedie(var.power=var.p,link.power=0),x=TRUE, data=d2)
#summary(modGlm)
fitted.v<-modGlm$fitted.values
residuals=dist-fitted.v
plot(residuals,xlab="Observation Number",ylab="Raw Residuals",
      main="Plot of the raw residuals, model 7")
abline(0,0) # add a horizontal line at 0
#-----model 8
var.power=tweedie.profile(dist~factor(emp_cat)+factor(paid_cat)+
                           factor(hhsize_group)+factor(wlth_index),
                           p.vec=seq(1.4,1.75,length=10),
                           do.ci=TRUE, method="saddlepoint", data = d2)
var.p=var.power$p.max
modGlm<-glm(dist~factor(emp_cat)+factor(paid_cat)+
             factor(hhsize_group)+factor(wlth_index),
```

```

        family=tweedie(var.power=var.p,link.power=0),x=TRUE, data=d2)
#summary(modGlm)
fitted.v<-modGlm$fitted.values
residuals=dist-fitted.v
plot(residuals,xlab="Observation Number",ylab="Raw Residuals",
      main="Plot of the raw residuals, model 8")
abline(0,0) # add a horizontal line at 0
#-----model 9
var.power=tweedie.profile(dist~factor(emp_cat)+factor(paid_cat)+
                          factor(hhsize_group)+factor(wlth_index)+
                          factor(school_cat),
                          p.vec=seq(1.4,1.75,length=10),
                          do.ci=TRUE, method="saddlepoint", data = d2)
var.p=var.power$p.max
modGlm<-glm(dist~factor(emp_cat)+factor(paid_cat)+
            factor(hhsize_group)+factor(wlth_index)+
            factor(school_cat),
            family=tweedie(var.power=var.p,link.power=0),x=TRUE, data=d2)
#summary(modGlm)
fitted.v<-modGlm$fitted.values
residuals=dist-fitted.v
plot(residuals,xlab="Observation Number",ylab="Raw Residuals",
      main="Plot of the raw residuals, model 9")
abline(0,0) # add a horizontal line at 0
#-----model 10
var.power=tweedie.profile(dist~factor(emp_cat)+factor(paid_cat)+
                          factor(hhsize_group)+factor(wlth_index)+
                          factor(school_cat),
                          p.vec=seq(1.4,1.75,length=10),
                          do.ci=TRUE, method="saddlepoint", data = d2)
var.p=var.power$p.max

```



```

modGlm<-glm(dist~factor(emp_cat)+factor(paid_cat)+
            factor(hhsize_group)+factor(wlth_index)+
            factor(school_cat)+factor(new_age_work_group),
            family=tweedie(var.power=var.p,link.power=0),x=TRUE, data=d2)
#summary(modGlm)
fitted.v<-modGlm$fitted.values
residuals=dist-fitted.v
plot(residuals,xlab="Observation Number",ylab="Raw Residuals",
     main="Plot of the raw residuals, model 10")
abline(0,0) # add a horizontal line at 0

```

R code for QQ normal Plot 4.6

```

par(mfrow = c(2, 2))
attach(d2)
#-----model 7
var.power=tweedie.profile(dist~factor(emp_cat)+factor(paid_cat)+
                          factor(hhsize_group),
                          p.vec=seq(1.4,1.75,length=10),
                          do.ci=TRUE, method="saddlepoint", data = d2)
var.p=var.power$p.max
modGlm<-glm(dist~factor(emp_cat)+factor(paid_cat)+
            factor(hhsize_group),
            family=tweedie(var.power=var.p,link.power=0),x=TRUE, data=d2)
#summary(modGlm)
fitted.v<-modGlm$fitted.values
residuals=dist-fitted.v
quantile=qres.tweedie(modGlm) # Quantile residuals
qqnorm(quantile, main = "Normal probability plot, model 7",
       xlab="Standard Normal Quantiles", ylab="Quantile Residuals")
qqline(quantile) # Normality line

```

```
#-----model 8
var.power=tweedie.profile(dist~factor(emp_cat)+factor(paid_cat)+
                          factor(hhsize_group)+factor(wlth_index),
                          p.vec=seq(1.4,1.75,length=10),
                          do.ci=TRUE, method="saddlepoint", data = d2)
var.p=var.power$p.max
modGlm<-glm(dist~factor(emp_cat)+factor(paid_cat)+
            factor(hhsize_group)+factor(wlth_index),
            family=tweedie(var.power=var.p,link.power=0),x=TRUE, data=d2)
#summary(modGlm)
fitted.v<-modGlm$fitted.values
residuals=dist-fitted.v
quantile=qres.tweedie(modGlm) # Quantile residuals
qqnorm(quantile, main = "Normal probability plot, model 8",
       xlab="Standard Normal Quantiles", ylab="Quantile Residuals")
qqline(quantile) # Normality line
#-----model 9
var.power=tweedie.profile(dist~factor(emp_cat)+factor(paid_cat)+
                          factor(hhsize_group)+factor(wlth_index)+
                          factor(school_cat),
                          p.vec=seq(1.4,1.75,length=10),
                          do.ci=TRUE, method="saddlepoint", data = d2)
var.p=var.power$p.max
modGlm<-glm(dist~factor(emp_cat)+factor(paid_cat)+
            factor(hhsize_group)+factor(wlth_index)+
            factor(school_cat),
            family=tweedie(var.power=var.p,link.power=0),x=TRUE, data=d2)
#summary(modGlm)
fitted.v<-modGlm$fitted.values
residuals=dist-fitted.v
quantile=qres.tweedie(modGlm) # Quantile residuals
```

```

qqnorm(quantile, main = "Normal probability plot, model 9",
       xlab="Standard Normal Quantiles", ylab="Quantile Residuals")
qqline(quantile) # Normality line
#-----model 10
var.power=tweedie.profile(dist~factor(emp_cat)+factor(paid_cat)+
                          factor(hhsize_group)+factor(wlth_index)+
                          factor(school_cat),
                          p.vec=seq(1.4,1.75,length=10),
                          do.ci=TRUE, method="saddlepoint", data = d2)
var.p=var.power$p.max
modGlm<-glm(dist~factor(emp_cat)+factor(paid_cat)+
            factor(hhsize_group)+factor(wlth_index)+
            factor(school_cat)+factor(new_age_work_group),
            family=tweedie(var.power=var.p,link.power=0),x=TRUE, data=d2)
#summary(modGlm)
fitted.v<-modGlm$fitted.values
residuals=dist-fitted.v
quantile=qres.tweedie(modGlm) # Quantile residuals
qqnorm(quantile, main = "Normal probability plot, model 10",
       xlab="Standard Normal Quantiles", ylab="Quantile Residuals")
qqline(quantile) # Normality line

```

R code for Raw residuals versus the Linear Predictor plot [4.7](#)

```

par(mfrow = c(2, 2))
attach(d2)
#-----model 7
var.power=tweedie.profile(dist~factor(emp_cat)+factor(paid_cat)+
                          factor(hhsize_group),
                          p.vec=seq(1.4,1.75,length=10),
                          do.ci=TRUE, method="saddlepoint", data = d2)

```

```

var.p=var.power$p.max
modGlm<-glm(dist~factor(emp_cat)+factor(paid_cat)+
            factor(hhsize_group),
            family=tweedie(var.power=var.p,link.power=0),x=TRUE, data=d2)
#summary(modGlm)
fitted.v<-modGlm$fitted.values
residuals=dist-fitted.v
plot(log(fitted.v),residuals,xlab="Linear Predictor", ylab="Pearson
      Residuals",main="Plot of raw residuals vs linear predictor, model 7")
#-----model 8
var.power=tweedie.profile(dist~factor(emp_cat)+factor(paid_cat)+
                          factor(hhsize_group)+factor(wlth_index),
                          p.vec=seq(1.4,1.75,length=10),
                          do.ci=TRUE, method="saddlepoint", data = d2)
var.p=var.power$p.max
modGlm<-glm(dist~factor(emp_cat)+factor(paid_cat)+
            factor(hhsize_group)+factor(wlth_index),
            family=tweedie(var.power=var.p,link.power=0),x=TRUE, data=d2)
#summary(modGlm)
fitted.v<-modGlm$fitted.values
residuals=dist-fitted.v
plot(log(fitted.v),residuals,xlab="Linear Predictor", ylab="Pearson
      Residuals",main="Plot of raw residuals vs linear predictor, model 8")
#-----model 9
var.power=tweedie.profile(dist~factor(emp_cat)+factor(paid_cat)+
                          factor(hhsize_group)+factor(wlth_index)+
                          factor(school_cat),
                          p.vec=seq(1.4,1.75,length=10),
                          do.ci=TRUE, method="saddlepoint", data = d2)
var.p=var.power$p.max
modGlm<-glm(dist~factor(emp_cat)+factor(paid_cat)+

```

```
        factor(hhsize_group)+factor(wlth_index)+
        factor(school_cat),
    family=tweedie(var.power=var.p,link.power=0),x=TRUE, data=d2)
#summary(modGlm)
fitted.v<-modGlm$fitted.values
residuals=dist-fitted.v
plot(log(fitted.v),residuals,xlab="Linear Predictor", ylab="Pearson
      Residuals",main="Plot of raw residuals vs linear predictor, model 9")
#-----model 10
var.power=tweedie.profile(dist~factor(emp_cat)+factor(paid_cat)+
        factor(hhsize_group)+factor(wlth_index)+
        factor(school_cat),
    p.vec=seq(1.4,1.75,length=10),
    do.ci=TRUE, method="saddlepoint", data = d2)
var.p=var.power$p.max
modGlm<-glm(dist~factor(emp_cat)+factor(paid_cat)+
        factor(hhsize_group)+factor(wlth_index)+
        factor(school_cat)+factor(new_age_work_group),
    family=tweedie(var.power=var.p,link.power=0),x=TRUE, data=d2)
#summary(modGlm)
fitted.v<-modGlm$fitted.values
residuals=dist-fitted.v
plot(log(fitted.v),residuals,xlab="Linear Predictor", ylab="Pearson
      Residuals",main="Plot of raw residuals vs linear predictor, model 10")
```

Appendix D: Supplementary Tables and Figures

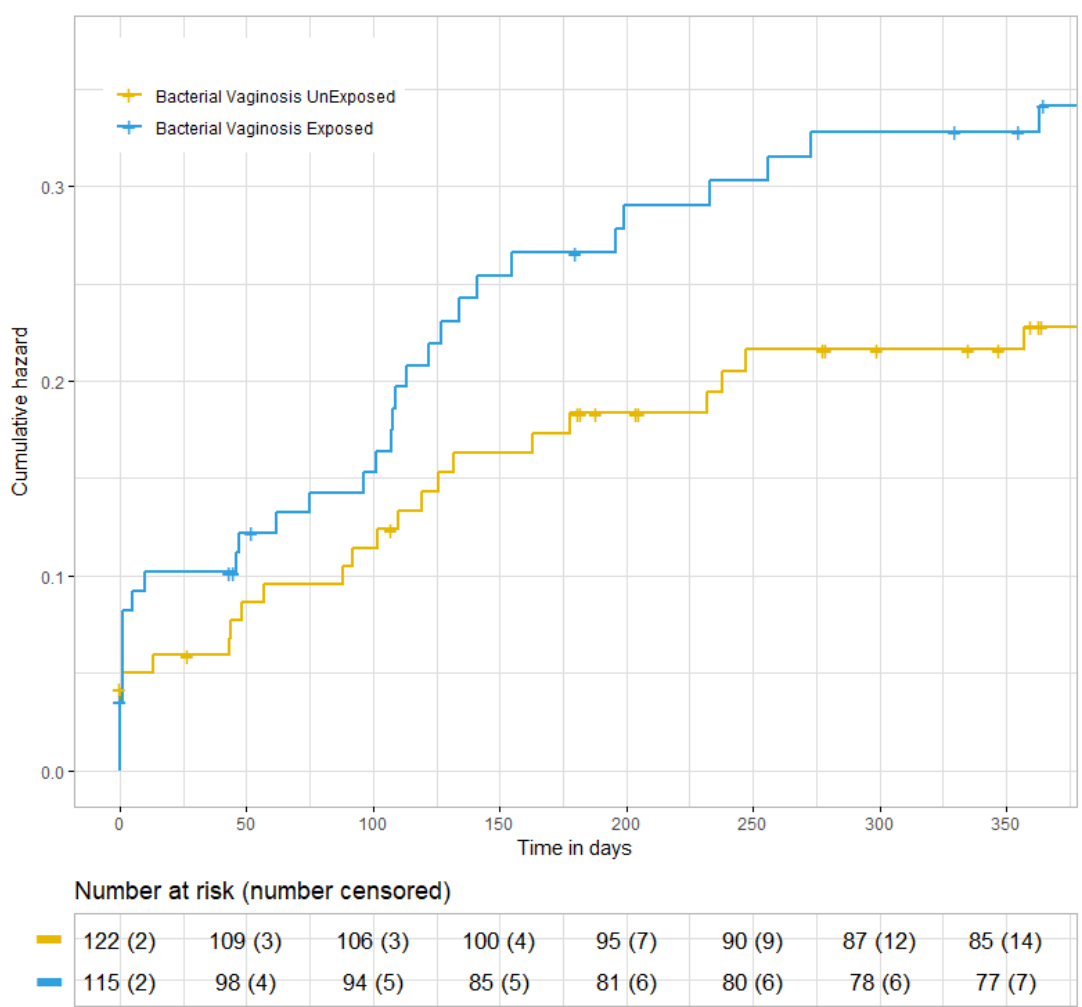


FIGURE 1: *Infant mortalities for one year between infants whose mothers are exposed to the BV and unexposed*

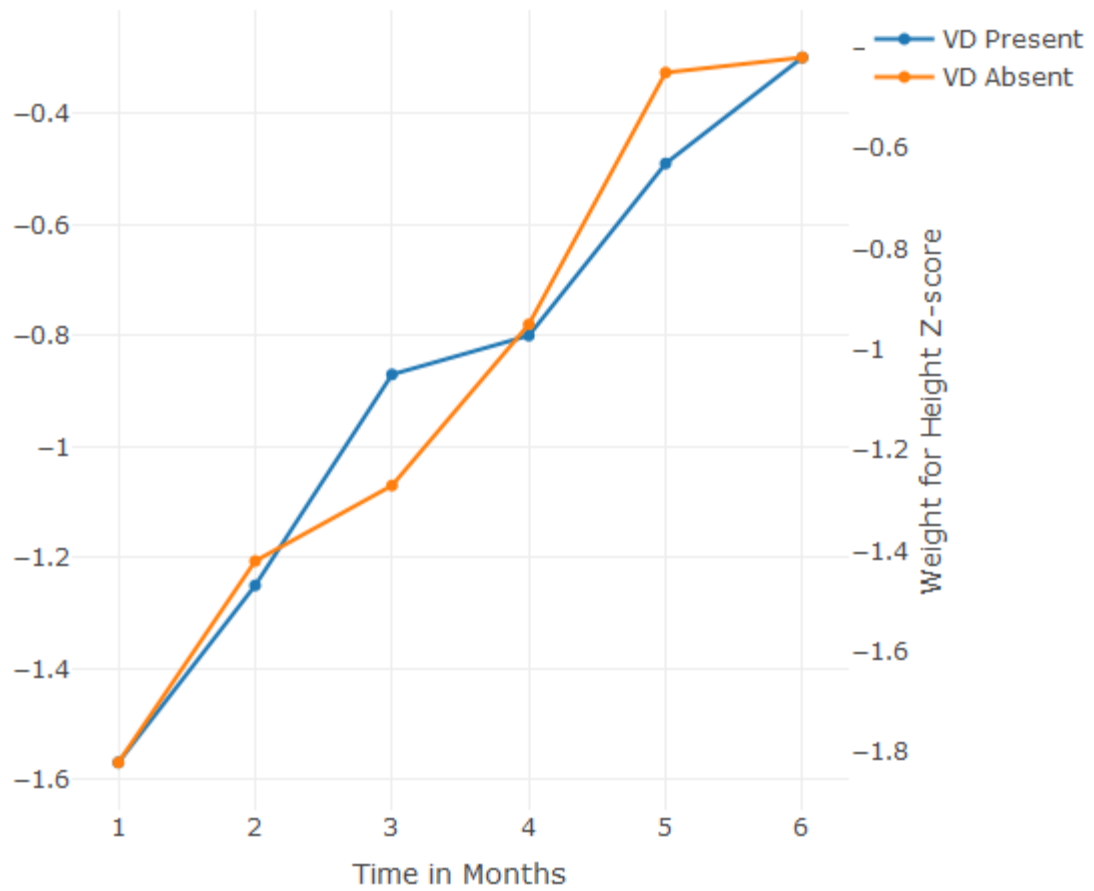


FIGURE 2: *Weight gain for age for the exposed and unexposed*

TABLE 2: *Compute the skewness parameter(α)*

Variable		Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
Time in Months		1.26	0.43	2.93	0.00	0.42	2.10
weight(in Kilograms)		-0.90	0.39	-2.28	0.02	-1.67	-0.13
Feeding(Reference Formula fed)		-0.32	0.92	-0.35	0.73	-2.11	1.48
Gender(Reference Male)		-2.46	0.97	-2.53	0.01	-4.36	-0.56
HIV status(Reference Positive)		1.88	1.52	1.24	0.22	-1.10	4.86
Vaginal Dysbiosis (Reference yes)		10.80	2.14	5.05	0.00	6.61	14.99
Time in Months							
	2	-0.46	0.95	-0.49	0.63	-2.32	1.39
	3	-0.77	1.09	-0.71	0.48	-2.91	1.37
	4	1.22	1.25	0.97	0.33	-1.24	3.68
	5	-0.53	1.29	-0.41	0.68	-3.05	2.00
Vaginal Dysbiosis *Time							
	1 2	-5.42	2.21	-2.45	0.01	-9.76	-1.08
	1 3	-5.60	2.50	-2.24	0.03	-10.49	-0.70
	1 4	-9.41	2.72	-3.46	0.00	-14.74	-4.08
	1 5	-9.18	2.42	-3.79	0.00	-13.93	-4.43
	1 6	-12.20	2.45	-4.98	0.00	-17.01	-7.40
(Intercept)		11.41	1.79	6.39	0.00	7.91	14.91
$\ln \alpha$		-2.56	0.05	-49.11	0.00	-2.66	-2.45
α		0.08	0.00			0.07	0.09

Appendix E: List of Publications and Conferences

1. N. Mwenda, R. Nduati, M. Kosgei and G. Kerich, "Infant Morbidities Differentials of HIV positive Mothers with Known Vaginal Dysbiosis Status: A Statistical Analysis with a Skewed Binary Outcome using Generalised Estimating Equations", accepted and presented (on demand poster) during the International AIDS 2020 conference, 6-11 July 2020, San Francisco and Oakland, USA.
2. N. Mwenda, R. Nduati, M. Kosgei and G. Kerich, "Morbidity incidence and Mortality in infants from HIV-1-infected mothers with bacterial Vaginosis in Kenya", accepted and presented (on demand poster) during the Center for HIV Identification, Prevention, and Treatment Services (CHIPTS) 2021 HIV Next Generation Conference, 22 January 2021, Los Angeles, CA.
3. N. Mwenda, R. Nduati, M. Kosgei and G. Kerich, "Effect of Bacterial Vaginosis (BV)-HIV-1 Co-Existence on Maternal and Infant health: A Secondary Data Analysis"
Frontiers in pediatrics, doi.org/10.3389/fped.2021.544192 *status published*
4. Mwenda N, Nduati R, Kosgei M, Kerich G (2021) Skewed logit model for analyzing correlated infant morbidity data.
PLoS ONE 16(2): e0246269. [doi:10.1371/journal.pone.0246269](https://doi.org/10.1371/journal.pone.0246269) *status published*
5. N. Mwenda, R. Nduati, M. Kosgei and G. Kerich, "Association of differences in household size, employment status, and amount paid for services with distance traveled for inpatient care in Kenya", accepted and presented (*on demand oral*

- presentation*) during Canadian Association for Health Services and Policy Research (CAHSPR) annual Conference, 19-21 May 2021, Canada
6. N. Mwenda, R. Nduati, M. Kosgei and G. Kerich, "Differences in Household size, Employment Status and Ability to pay for the service, are Associated with Distance Traveled for Inpatient Care in Kenya" Submitted to PLOS ONE publications, *status minor revisions*
 7. N. Mwenda, R. Nduati, M. Kosgei and G. Kerich, "Predictors of Household spending on Out-patient Expenses in Kenya", accepted and orally presented at the 6th Strathmore International mathematics Conference (SIMC 2021), held from June 28th to July 2nd 2021.
 8. N. Mwenda, R. Nduati, M. Kosgei and G. Kerich, "Predictors of Household spending on Out-patient Expenses in Kenya" Submitted to *frontiers in health Economics*, *status minor revisions*