Statistical modeling of acute HIV infection from a cohort of high-risk individuals in South Africa

UNIVERSITY OF
KWAZULU - NATAL

INYUVESI
YAKWAZULU-NATALI

Ashenafi Argaw Yirga

July, 2022

Statistical modeling of acute HIV infection from a cohort of high-risk individuals in South Africa

by

Ashenafi Argaw Yirga

A thesis submitted to the

University of KwaZulu-Natal

in fulfilment of the academic requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

APPLIED STATISTICS

UNIVERSITY OF
KWAZULU - NATAL
INYUVESI
YAKWAZULU-NATALI

COLLEGE OF AGRICULTURE, ENGINEERING AND SCIENCE

SCHOOL OF MATHEMATICS, STATISTICS AND COMPUTER SCIENCE

PIETERMARITZBURG, SOUTH AFRICA.

Declaration

I, Ashenafi Argaw Yirga, declare that this dissertation titled "Statistical modeling of acute HIV infection from a cohort of high-risk individuals in South Africa" and the work presented in it is my original work and has not been submitted for any degree or examination at any other university, does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons, and does not contain text, graphics or tables copied and pasted from the internet, unless specifically acknowledged, and the source being detailed in the thesis and in the reference sections.

The research work described in this dissertation was carried out under the supervision and direction of Prof. Sileshi Fanta Melesse and co-supervised by Prof. Henry Godwell Mwambi and Dr. Dawit Getnet Ayele as a Ph.D. programme of study at the University of KwaZulu-Natal (UKZN).

Pietermaritzburg, July 2022

[redacted]	11/07/2022
Ashenafi Argaw Yirga	Date
[redacted]	11/07/2022
Prof. Sileshi Fanta Melesse	Date
[redacted]	12/07/2022
Prof. Henry Godwell Mwambi	Date
[redacted]	11/07/2022
Dr. Dawit Getnet Ayele	Date

Acknowledgements

"Behold, Michael, one of the chief angel-princes, came to help me"
— Daniel 10:13

First and foremost, I would like to praise God Almighty for all the blessings granted unto me and for His help which has been with me. I give thanks to the blessed, full of glory, holy and pure Virgin Mary, our Lady, Mother of God, and my guardian archangel Saint Michael.

All through my many years as a student—from grade school and high school in Jijiga city and Butajira city to my time as an undergraduate at Addis Ababa University, Ethiopia, to honors and postgraduate school at the University of KwaZulu-Natal—I have had the great fortune of being taught by many quality educators, doctors, and professors. I am grateful to all of them. Especially in my journey towards this degree, I have found mentors, role models, and pillars of supporters in my guide, **Professor Sileshi Melesse**, **Professor Henry Mwambi**, and **Doctor Dawit Ayele**. They have been there for me, always providing their heartfelt support and guidance. They have given me motivation and invaluable suggestions in earning this degree and increased my understanding of Statistics. They have given me all the freedom to pursue my research while ensuring that I stayed on course and did not deviate from my research core. I shall eternally be grateful to them for their assistance and for taking valuable time out of their busy schedules to improve my thesis and the published papers; without their able guidance, this dissertation and achievement would not have been possible. Their spirit of research has had a significant and life-long positive impact on my professional development. It has been a great honor and privilege working with them.

Prof. Sileshi, my major supervisor, I would like to express my deep and sincere gratitude to you for accepting me as a doctoral candidate and given me the opportunity to work with you. I cannot thank you enough for your academic and professional guidance, patience, encouragement, motivation, advice, and prompt reviews of my thesis, which made my Ph.D. experience productive and stimulating. **Prof. Mwambi**, I am deeply indebted to your continuous support, essential inputs, valu-

able corrections, constructive ideas, and quick responses to my quest, which have been great contributors to the completion of this thesis. Thank you for introducing me to Dr. Nonhlanhla Yende-Zuma and facilitating to get the study data from CAPRISA. I am also grateful for the financial support that I have received during my time as a post-graduate student and for selecting me to receive the DELTAS SACAB fellowship, which allowed me to focus on my research work and writing my thesis. I want to express special and sincere gratitude to **Dr. Dawit**, whose kind support, patience, and guidance since I started studying at UKZN is tremendous. His erudition, enthusiasm, and personality are motivational to me and firmly anchored me through the very end. He spared his valuable time whenever I approached him and showing me the way ahead. I will always remain thankful to him.

I want to thank all staff members in the UKZN, friends, colleagues, and office mates for the nice working environment and our fruitful discussions and sharing of information. I am also gratefully acknowledged CAPRISA for giving me access to one of their world-class CAPRISA 002: Acute Infection study data for my doctoral study work, without which this work may not have seen the light of day. Many thanks to Dr. Nonhlanhla Yende-Zuma (Head of Biostatistics and Data Management at CAPRISA) for her kindness, cooperation, assistance, and technical support. CAPRISA is funded by the National Institute of Allergy and Infectious Diseases (NIAID), the National Institutes for Health (NIH), and the U.S. Department of Health and Human Services (grant: AI51794).

I thank the DELTAS Africa Initiative fund for funding both my doctoral and master's degree studies. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [grant 107754/Z/15/Z], DELTAS Africa Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB) programme and the UK government.

I dedicate this work to my beloved, kind, and Godly late mother, **W/ro Tsedash Ameme Woldesillase**, who passed away so suddenly in October 2020. Her dreams for me have resulted in this achievement, and without her prayer, love, blessings, nurturing, and upbringing, I would not have been where I am today and what I am today. Had it not been for my mom's unflinching insistence and support, my dreams of excelling in education would have remained mere dreams–the desire to make my mom pleased propelled me to achieve this doctoral degree. I thank my mom with all my heart. 2020 AD was really a challenging year for me. I want to extend my ac-

knowledgment to my humble late grandfather **Mr. Yirga Dibaba Jabora**, who lived more than 100 years and rested in God's love in November 2020. I am thankful to receive more blessings, love, and kindness from my grandpa. I am fortunate to have had a good time and loving memory with my grandpa. May God bless their soul as they ascend to heaven with Him.

I feel very thankful to my sisters and brother for putting up with me in difficult times where I felt stumped and for goading me onto follow my dream of getting this degree. This would not have been possible without their unconditional love, continuous support, and encouragement always given to me. This achievement would not have also been possible without the support, inspiration, and guidance from several people, and I would like to take this opportunity to express my gratitude to everyone who, in one way or another, contributed to the successful accomplishment of my doctoral degree.

Mr. Ashenafi Argaw Yirga
Pietermartizburg, July 2022

Vita

2011 ..... B.Sc., Statistics,
Addis Ababa University.

2016 ..... B.Sc., Honours Statistics,
University of KwaZulu-Natal.

Yirga (2018) ..... M.Sc., Statistics (*Cum Laude*),
University of KwaZulu-Natal.

Abstract

In this dissertation, longitudinal data modeling approaches to analyze data on CD4 cell counts measured repeatedly in HIV-infected patients enrolled in the Centre for the AIDS Programme of Research in South Africa are investigated. Longitudinal data, or repeated measurements data, is a specific form of multilevel data. In longitudinal studies, repeated observations are made on an individual on one or more outcomes, including covariates information at a baseline and over time. Mixed-effects models have become popular for modeling longitudinal data. This statistical procedure also permits the estimation of variability in hierarchically structured data and examines the impacts of factors at different levels. Since longitudinal studies are often faced with the incompleteness of the data due to partially observed subjects, the mixed-effects model is by its very nature able to deal with unbalanced data of this nature. Therefore, the study adopts the mixed-effects model and identifies whether specific clinical and sociodemographic factors present in the data influenced CD4 count in a cohort of HIV-infected patients.

Since it is of great interest for a biomedical analyst or an investigator to correctly model the CD4 cell count or disease biomarkers of a patient in the presence of covariates or factors determining the disease progression over time, the Poisson regression approach, which explain variability in counts, is considered. The Poisson generalized mixed-effects models can be an appropriate choice for repeated count data. However, this model is not realistic because of the restriction that the mean and variance are equal. Therefore, the Poisson mixed-effects model is replaced by the negative binomial mixed-effects model. The later model effectively managed over-dispersion of the longitudinal data. We evaluate and compare the proposed models and their application to model CD4 cell counts of HIV-infected patients recruited in the study data set. The results reveal that the negative binomial mixed-effects model has appropriate properties and outperforms the Poisson mixed-effects model in terms of handling the over-dispersion of the data. Multiple imputation techniques are also used to handle missing values in the dataset to validate parameter estimates in modeling the negative binomial mixed-effects model by assuming a missing at random missingness.

To illustrate the full conditional distribution of the repeated outcome, a quantile mixed-effects model is employed. This gives greater inclusive statistical modeling than conventional ordinary mixed models. Quantile regression offers an invaluable tool to discern effects that would be missed by other conventional regression models, which are solely based on modeling conditional mean. The quantile regression model that assumes asymmetric Laplace distribution for the error term was applied to longitudinal CD4 count data. The exact maximum likelihood estimation of the covariate effects and variance-covariance elements in the quantile mixed-effects model was implemented using the Stochastic Approximation Expectation-Maximization algorithm. In the model, multiple random effects are also incorporated to consider the correlation among the observations. Thus, we obtain robust parameter estimates for various conditional distribution positions that communicate an inclusive and more complete picture of the effects.

Furthermore, to get more insights into the functional relationship between the response variable and the covariates, the generalized additive mixed-effects models, such as the additive negative binomial mixed-effects model, a versatile model used to better understand and analyze complex nonlinear trajectories in an overdispersed longitudinal data, is applied. Following the additive negative binomial mixed-effects model, an attempt to fit additive quantile mixed-effects model, an efficient and flexible framework for nonparametric as well as parametric longitudinal forms of data analysis focused on features of the outcome beyond its central tendency, was made.

The response variable at hand is a CD4 count of HIV-infected patients as a function of Highly Active Antiretroviral Therapy initiation and other relevant baseline characteristics of the patients. Thus, even though this is a biostatistics methodological dissertation research, some interesting clinical and sociodemographic findings are also discussed. Discussion and conclusion of the results from the proposed models with a suggestion of possible further research avenues completed the study.

List of Publications

This thesis is based on the following publications and report:

- Yirga et al. (2020a). Modelling CD4 counts before and after HAART for HIV infected patients in KwaZulu-Natal South Africa. ***African Health Sciences***, 20(4), pp.1546-61.
- Yirga et al. (2020b). Negative binomial mixed models for analyzing longitudinal CD4 count data. ***Scientific Reports***, 10(1), pp.1-15.
- Yirga et al. (2021a). Additive quantile mixed effects modelling with application to longitudinal CD4 count data. ***Scientific Reports***, 11(1), pp.1-12.
- Yirga et al. (2021b). Analyzing longitudinal CD4 count of HIV-infected patients using generalized additive mixed-effects model. [Under Review].
- Yirga et al. (2022). Application of quantile mixed-effects model in modeling CD4 count from HIV-infected patients in KwaZulu-Natal South Africa. ***BMC Infectious Diseases***, 22(1), pp.1-11.

Contents

List of Figures

List of Tables

Abbreviations and Notation

AI	Acute Infection
AIC	Akaike Information Criterion
AICC	Corrected AIC
AIDS	Acquired Immunodeficiency Syndrome
ALD	Asymmetric Laplace Distribution
AM	Additive Model
AMM	Additive Mixed-Effects Model
AQM	Additive Quantile Model
AQMM	Additive Quantile Mixed-Effects Model
ART	Antiretroviral Therapy
ARV	Antiretroviral (*drug*)
BIC	Bayesian Information Criterion
BMI	Body Mass Index
BLUE	Best Linear Unbiased Estimator
BMI	Body Mass Index
CAIC	Consistent Akaike's Information Criterion
CAPRISA	Centre for the AIDS Programme of Research in South Africa
CD4	Cluster of Difference 4 cell (*T-lymphocyte cell*)
CDC	United States Centers for Disease Control and Prevention
cdf	Cumulative Distribution Function
CI	Confidence Interval
E-step	Expectation step
EM	Expectation-Maximization
GAM	Generalized Additive Model
GAMLESS	Generalized Additive Model for Location, Scale, and Shape
GAMM	Generalized Additive Mixed-Effects Model
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed-Effects Model

HAART	Highly Active Antiretroviral Therapy
HIV	Human Immunodeficiency Virus
HQIC	Hannan-Quinn Information Criterion
IC	Information Criterion
iid	Independent and Identically Distributed
LP	Linear Programming
LMM	Linear Mixed-Effects Model
M-step	Maximization step
MI	Multiple Imputations
ML	Maximum Likelihood
NBMM	Negative Binomial Mixed-Effects Model
pdf	Probability Density Function
PMM	Poisson Mixed-Effects Model
QR	Quantile Regression
QR-LMM	Quantile Regression for Linear Mixed-Effects Models
R	R Statistical Software
REML	Restricted/*Residual* Maximum Likelihood
r.v.	Random Variable
SAEM	Stochastic Approximation version of the EM algorithm
SAS	Statistical Analysis System
SE	Standard Error
SMM	Semiparametric Mixed-Effects Model
SPSS	Statistical Package for Social Science
Stata	Statistical Software (Statistics and data)
STD	Sexually Transmitted Disease
UN	Unstructured covariance
UNAIDS	Joint United Nations Programme on HIV and AIDS
UNDP	United Nations Development Programme
UNESCO	United Nations Educational, Scientific and Cultural Organization
UNICEF	United Nations Children's Fund
VL	Viral Load refers to the number of HIV copies per cubic millimeter
WHO	World Health Organization

$\boldsymbol{A}'$	The transpose of a matrix $\boldsymbol{A}$
$\|\boldsymbol{A}\|$	The determinant of a square matrix $\boldsymbol{A}$
$\boldsymbol{A}^{-1}$	The inverse of a square, non-singular matrix $\boldsymbol{A}$
$\lceil a \rceil$	The *ceiling* function, which is the smallest integer $\geq a$
ALD(μ, σ, τ)	ALD law with mean, variance, and mode at τ, respectively
$trace(\boldsymbol{A})$	The trace of the square matrix $\boldsymbol{A}$
$rank(\boldsymbol{A})$	The rank of the matrix $\boldsymbol{A}$
d	The differential operator
$\boldsymbol{I}$	Identity matrix
I(θ)	The Fisher information about θ
$I(\cdot)$	An indicator function
log	Natural logarithm
$\min f(\cdot)$	The minimum value of the function f
$\arg\min f(\cdot)$	The value of the argument of f that yields its minimum value
$\mathbb{N}$	The set of natural numbers
$N(\mu, \sigma^2)$	Normal distribution with mean μ, and variance σ^2
$P_r(A)$	The probability of the event A
$\mathbb{R}$	The field of real numbers
$\mathbb{R}^d$	d-dimensional Euclidean space with elements in $\mathbb{R}$
$\mathbb{R}_+^d$	d-dimensional Euclidean space with positive elements in $\mathbb{R}$
x_i	The i^{th} element of a sample
$x_{(i)}$	The i^{th} order statistic
$x^{(i)}$	The value of x at the i^{th} iteration
$\|x\|$	The absolute value of x
$\bar{x}$	The mean of a sample
χ^2	Chi-square distribution
$\Gamma(\alpha)$	The gamma function, $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx = (\alpha-1)!$
$\|\boldsymbol{a}\|_{\boldsymbol{G}}$	The norm of a vector $\boldsymbol{a}$ with respect to a matrix $\boldsymbol{G}$ which is the same as $\sqrt{(a^T G a)}$
0_n	$n \times 1$ vector of zeros
1_n	$n \times 1$ vector of ones
$\bigoplus$	Kronker sum of matrices (direct sum)

DEDICATED TO MY BELOVED DEEPLY MISSED MOM,
W/ro TSEDASH AMEME WOLDE-SILLASE
(***TSED***).

Chapter 1

Introduction

Longitudinal studies are characterized as studies in which the response variable is measured in the same individual on several different occasions. In longitudinal studies, the observations of one individual over time are not independent of each other. Thus it is vital to apply special statistical techniques, which consider the fact that the repeated observations of each individual are correlated. In contrast, cross-sectional data refers to the situation at one particular point in time. The main advantage of a longitudinal study compared to a cross-sectional study is that it can study the individual development of a specific outcome variable over time. In addition to this, the individual development of a particular outcome variable can be related to the individual development of other variables. For example, HIV patients may be followed over time, and monthly measures of disease biomarkers such as CD4 counts and viral load are collected to characterize immune status and disease burden, respectively. Such repeated measures data require special statistical techniques for accurate analysis and inference.

Longitudinal data analysis is widely used for at least three reasons: to increase the sensitivity by making within-subject comparisons, to study changes over time, and to use subjects efficiently once they are enrolled in a study (Twisk, 2013; Hedeker & Gibbons, 2006; Der & Everitt, 2012). Repeated measurements can compensate for small sample sizes because an individual is observed more than once compared to a cross-sectional study. The covariance structure of the observed data makes longitudinal data analysis distinct. For the analysis to be valid, one must appropriately model the covariance among repeated measures. Although the covariance structure is not the prime interest of the study, it is essential for valid inference (Kincaid, 2005; Kowalchuk et al., 2004). Therefore, a lot of efforts are needed at the beginning of the statistical analysis to assess the covariance structure of the data. Traditional methods

for longitudinal data such as Analysis of Variance (ANOVA) and Multivariate Analysis of Variance (MANOVA) are of limited use because of the restrictive assumptions concerning the variance-covariance structure of the repeated measures (Liu, 2015).

For this reason, mixed-effects models have become famous for modeling longitudinal data. This statistical procedure also permits the estimation of variability in hierarchically structured data and examines the impacts of factors at distinctive levels (Taris, 2000; Brown & Prescott, 2014; Yirga et al., 2020a). Since longitudinal studies are often faced with the incompleteness of the data due to partially observed subjects, the mixed-effects model is by its very nature able to deal with unbalanced data of this nature (Yirga et al., 2020a). There are also several methods suited to dealing with longitudinal data, such as generalized estimating equations and generalized linear mixed-effects models (Diggle et al., 2002).

Count data are ubiquitous in epidemiological studies. This sort of data assumes only non-negative integer values (i.e. 0, 1, 2, . . .). The most commonly used model for count data is the Poisson distribution and its related enhancement, such as the Poisson-gamma model, to account for overdispersion and heterogeneity (Brown & Prescott, 2014; Diggle et al., 2002; Weiss, 2005; Molenberghs et al., 2010). However, these approaches confine the analysis of differences among units in terms of the mean of the dependent variable, and they employ parametric models based on the distributional hypothesis (Davino et al., 2013). Further, in some cases, it might be challenging to find a suitable transformation to normalize the outcome, or some resistance to outliers may be desired. An effective solution to all these issues is given by focusing on the conditional quantiles of the longitudinal outcome (Koenker, 2005b). The conditional quantile regression method, which measures the complete conditional distribution of the outcome variable, was developed to assess covariate effects at any subjective quantiles of the outcome. In addition, quantile regression methods do not impose any distribution assumption on the error, except that the error term has a zero conditional quantile, such as the asymmetric Laplace distribution (ALD) (Wichitaksorn et al., 2014; Galarza et al., 2017).

The generalized additive mixed-effects models, such as the additive negative binomial mixed-effects model, can be used to better understand and analyze complex nonlinear trajectories and get more insights into the functional relationship between the outcome and the covariates, especially for over-dispersed longitudinal data. A recently developed model, additive quantile mixed-effects model, also offers an efficient and flexible framework for nonlinear and linear longitudinal forms of data analysis focused on features of the outcome beyond its central tendency. The practi-

cal motivation of these methods is applied to the CD4 count of HIV-infected patients from the Centre for the AIDS Programme of Research in South Africa (CAPRISA) 002 Acute Infection Study dataset.

1.1 Motivational Background

After it was identified by scientists as the human immunodeficiency virus (HIV) and the cause of acquired immunodeficiency syndrome (AIDS) in 1983, HIV has spread persistently, triggering one of the most severe pandemics ever documented in human history. More than 77 million individuals have been infected with HIV. More than 34 million individuals have died due to AIDS-related causes worldwide since the pandemic started, and 7000 new infections were reported daily, according to the UNAIDS report in 2019 (Yirga et al., 2020b). Globally, 36.9 million [31.1-43.9 million] people lived with HIV at the end of 2017. An estimated 0.8% [0.6-0.9%] of adults aged 15-49 years worldwide live with HIV, even though the burden of the epidemic continues to differ considerably between countries (Geneva, 2017; UNAIDS, 2019). However, intensive global efforts to battle the pandemic is making a significant difference. Despite recent advancements in HIV prevention, care, and treatment, which have modestly decreased the total number of new infections and deaths every year, AIDS and AIDS-related illnesses are still among the driving causes of loss of life globally (Yirga et al., 2020b). The impacts of HIV are far-reaching, which include reduced life expectancy, decreased economic development, and increased health costs. These consequences subsequently damage social and political cohesion and block the progression of worldwide health objectives-posing a risk to countrywide security and the steadiness of numerous nation-states (Geneva, 2017; UNAIDS, 2017, 2019; Yirga et al., 2020b).

HIV/AIDS has been a characterizing challenge of our time. Studies show that Africa carries the most burden for HIV/AIDS compared to other continents in the world. HIV/AIDS has been researched, written about, and discussed numerous times. This shows that HIV/AIDS remains a critical worldwide issue and obstruction to advancement. There are numerous articles published dealing with an assortment of viewpoints of HIV/AIDS. Even though the HIV epidemic appears to be established in multiple nations, this is not steady over globally. Numerous nations show an increment in incidence in recent years, including developed countries. Sub-Saharan Africa and southern Africa, in specific, are right now the region most affected by HIV/AIDS in the world (WHO et al., 2008; WHO, 2010; Yirga et al., 2020b). South Africa, found within the epicenter of the global scourge and with its disputable history of HIV healthcare policy, remains an essential region for attempting to get

the numerous social, psychological, political, and therapeutic components that play a role in the control of HIV/AIDS. Since South Africa is at the epicenter of the HIV/AIDS epidemic, lessons learned in South Africa are lessons for the general community (WHO et al., 2008; Shisana et al., 2014; Yirga et al., 2020b). Whereas South Africa may be one of the foremost affected by HIV globally, other regions in the world, such as nations in the south and southeast Asia and Latin America, are challenged with increasing prevalence rates and, in few cases, developing outbreaks. Furthermore, prevalence continues to increase across Eastern Europe, Central and East Asia, the Middle East, and North Africa (WHO et al., 2008; Yirga et al., 2020a,b).

The need for good and better health care is one of each human being's fundamental rights without qualification of race, religion, gender, political conviction, financial, or social condition. Women's health includes their emotional, social, and physical welfare and is determined by these factors and the economic setting of their lives, as well as by biology. However, health issues evade the longer part of women. In national and universal forums, women have emphasized that equality, the sharing of family duties, development, and peace are necessary conditions to achieve good health all through the life cycle. A major obstruction for women to the accomplishment of the most exceptional plausible standard of their health is an imbalance, both among people and among women in various geographical regions, social classes, and innate and ethnic bunches. In national and universal forums, women have emphasized that to achieve ideal well-being all through the life cycle, equality, together with the sharing of family duties, development, and peace, are necessary conditions. Women are biologically and socially more vulnerable to HIV infection, especially in developing countries (WHO et al., 2007; WHO, 2010; UN, 2014; amfAR, 2015).

HIV/AIDS and other sexually transmitted diseases (STD) have a devastating effect on women's health, mostly young ladies. The consequences of HIV/AIDS go beyond women's health to their part as mothers and caregivers and include their families' economic support and livelihoods. Thus, the social, development, and health consequences of HIV/AIDS and other sexually transmitted diseases have strong gender dimensions that cannot be ignored (Whelan, 1999; UN, 2014; amfAR, 2015). It needs to be emphasized that, except for those issues that are sex-specific, treatment algorithms for HIV-infected women do not contrast from those of men. Understanding the changing epidemiology of HIV using statistical disease models will allow the clinician to decide who may be at high risk and clarify the application of rules to avoid sequential HIV transmission. Although antiretroviral (ARV) recommendation presently remains the same for all individuals living with HIV, examining the progression of CD4 count or evolution of the viral load using data-driven models

will allow the clinician to interpret potential information accurately and cope with misdirection or distortion of the information due to patient-specific effects (Kassutto & Rosenberg, 2004; Cohen et al., 2011; Rosenberg et al., 2000).

CD4 cell count levels signify the well-being of an individual immune system (body's natural defense system against pathogens, infections, and illnesses). It also provides information about disease progression. CD4 cells are white blood cells (in a cubic millimeter of blood) that play an essential role in the immune system. A higher number shows a stronger immune system. The CD4 cell counts of a person who does not have HIV can be between 500 and 1500 per cubic millimeter (Hughson, 2017). Individuals living with HIV who have a CD4 count over 500 but whose immune response is still strong are usually in good health. However, individuals living with HIV who have a CD4 count below 200 are at high risk of developing severe illnesses and death (Hughson, 2017; Yirga et al., 2020b). With the CD4 count at deficient levels, patients' immunity is weak. If HIV-infected patients are not on treatment or not virally suppressed, they become vulnerable to acquire opportunistic infections (OIs) such as making them at risk of the new and ongoing coronavirus disease 2019 (COVID-19) infection, underlying illness and many others (Yirga et al., 2020b). The best strategy to avoid these infections and diseases is by enhancing the immune function level through HAART, a combination of multiple antiretroviral (ARV) drugs. HAART's fundamental goal is to prolong or stop the progression to AIDS and loss of life for those infected with HIV by suppressing and preventing the virus from making copies of itself. When the virus's level (viral load) in the blood is low or undetectable, there is less damage to the body's immune system and fewer HIV infection complications. Even though HIV treatment is prescribed for all individuals living with HIV, it is particularly critical for patients with low CD4 count to start treatment sooner rather than later and adhere to the treatment schedule (Yirga et al., 2020a,b). While researchers believe that early diagnosis and effective treatment are essential to effective control, more research is needed to understand better the adaptive, innate, and host responses that alter viral load set-point and consequently prognosis and infectiousness (Yirga et al., 2020a,b).

1.2 Objectives

The main objective of this thesis is to look for the appropriate statistical approaches that can help understand CD4 count progression and identify the potential risk factors affecting the CD4 count progression in HIV-infected individuals based on longitudinal observational data from the Centre for the AIDS Programme of Research in South Africa (CAPRISA) 002 Acute Infection Study. The primary outcome of interest

is CD4 count, which is widely used as a biomarker of HIV progression. The study's findings may suggest valuable insights that help further understand evolution of patients' CD4 cells and factors associated with it. Additionally, the study may improve understanding of patients' baseline characteristics that alter CD4 count progression and, consequently, respond to the treatment and improve their health. Most importantly, the research findings will contribute towards developing of intervention strategies (at the individual and population level) at the early stages of the disease.

1.3 Importance of the study

To introduce the most advanced level of care for people with HIV/AIDS in the health-care system, scale up the AIDS clinical treatment programs is an important measure. After introducing the AIDS clinical treatment programs, it is essential to monitor and counsel patients. This process optimizes the benefit of the medication. Furthermore, for most of the African countries, the costs of treatment programs are enormous. Because of this strict AIDS, clinical treatment medications should enforce to optimize benefits. Therefore, the study's outcome will not only provide clinicians with the factors associated with AIDS clinical treatments, but it will also offer within-patient differences in AIDS clinical treatment levels over time. That is, understanding specific barriers to medication adherence of individual patients will be valuable in the development and implementation of evidence-based interventions targeted at individual patients with poor adherence. The results should also be able to provide appropriate statistical models that can be useful to analyze AIDS clinical treatment data by governmental and non-governmental organizations to monitor adherence levels over time. In general, after identifying a good-fitting, realistic model, it can be used to project the short-term future of the HIV/AIDS epidemic, assuming that all parameter values and conditions remain constant. Statistical models can separate or disentangle the difference between individual-specific and population effects in the disease's evolution.

1.4 Outline of the dissertation

Five research papers have been produced from this thesis as previously stated under the "List of Publications and Reports" section. Four of these research papers (Yirga et al., 2020a,b, 2021a, 2022) have been published, and the rest one is currently under review for publication. Therefore, based on these research manuscripts with further details, the thesis's remainder is organized as follows: Chapter 1 gives some background about longitudinal studies, HIV/AIDS, disease biomarkers (CD4 cell count), the objective and importance of the study, and some current information on the his-

tory of women and HIV/AIDS by reviewing research materials that have been done in this field. Exploratory data analysis of the data set used for this study has also been done in all chapters.

Chapter 2 presents a mixed-effects model to analyze data on CD4 cell count repeatedly measured in HIV-infected patients enrolled in a subset of the CAPRISA study. Mixed-effects modeling is an advanced and vital method in statistics. It is a well-known method; therefore, we summarized the key aspects of the model relevant to this chapter's study. Using mixed-effects models for longitudinal data analysis helps to correctly account for the correlation of observations within a subject and quantify the heterogeneity between subjects due to unobserved factors. In addition, since longitudinal studies are often faced with the incompleteness of the data due to partially observed subjects, the mixed-effects model is by its very nature able to deal with unbalanced data of this nature. Thus, in Chapter 2, we adopt the mixed-effects model with appropriate random effects incorporated, including a flexible variance-covariance structure that gives the best fit and assesses the impact of HAART initiation and other relevant factors on the average CD4 count. We also studied additional works such as spatial covariance structure to account for spatial variability and overall influence diagnostics for the mixed effects and covariance parameters. Chapter 2 description is based on published work Yirga et al. (2020a).

In Chapter 3, the thesis presents a comparative study of Poisson regression and negative binomial in the context of generalized linear mixed-effects models to correctly model the CD4 count of a patient in the presence of factors determining the disease progression over time. The Poisson mixed-effects model can be an appropriate choice for repeated count data. However, this model is not realistic because of the equality restriction of the mean and variance. Therefore, it is replaced by the negative binomial mixed-effects model. The later model effectively manages the over-dispersion of the repeated count data. Evaluation and comparison of these models, as well as their application to a subset of CAPRISA data, are conducted. Chapter 3 describes the research paper published in Yirga et al. (2020b).

Chapter 4 presents a review of quantile regression and its mixed-effects extension for longitudinal data analysis. In this chapter, the longitudinal data's various conditional distribution is illustrated by employing the quantile regression mixed-effects model. It offers more rigorous and comprehensive estimates in contrast to the mean-based mixed-effects models, particularly in the case of skewed data. Thus, robust parameter estimates for various positions of the conditional distribution that communicates an inclusive and more complete picture of the effects are obtained. The

application to a subset of CAPRISA data is conducted. This chapter discusses the research paper published in Yirga et al. (2022).

Chapter 5 presents a versatile model, generalized additive mixed-effects models, to better understand and analyze complex nonlinear trajectories in longitudinal data. It helped us to combine linear and nonlinear terms in the model. In this chapter, we studied an additive negative binomial mixed-effects model to analyze the longitudinal CD4 count of HIV-infected patients in KwaZulu-Natal, South Africa, as a function of age, baseline BMI, and time non-parametrically as well as some covariates at hand parametrically. The results of the analysis give us more insights into the functional relationship between the response variable and the covariates. We illustrated the application to the CAPRISA 002 Acute Infection Study data. Chapter 5 describes the research paper presented in Yirga et al. (2021b).

In Chapter 6, we studied the additive quantile mixed-effects model, a recently developed model that has gained a great deal of popularity because it offers an efficient and flexible framework for nonlinear and linear longitudinal forms of data analysis focused on features of the outcome beyond its central tendency. We showed that additive quantile mixed-effects model could be used to obtain robust results, not only at the central location of the longitudinal outcome that may not be the best location to characterize the data but also at different locations of the conditional distribution that communicates an inclusive and more complete picture of the parametric as well as the nonparametric covariate effects. Chapter 6 discusses the research paper published in Yirga et al. (2021a).

Finally, Chapter 7 summarizes the preceding chapters' findings, discusses the implications of these findings, and suggests future research avenues. Appendices hold: additional results, the SAS- and R-codes to implement the results presented in Chapters 2 - 6, and supplementary materials.

Chapter 2

Modelling CD4 counts before and after HAART for HIV-infected patients in KwaZulu-Natal South Africa

2.1 Introduction

Multilevel data modeling allows us to account for the correlation of measurements and includes variables measured at different levels as well as model the variation at different levels. Longitudinal data, or repeated measurements data, is a specific form of multilevel data (Yirga et al., 2020a). In longitudinal studies, repeated observations are made on an individual on one or more outcomes, including covariate information at a baseline and over time. Measurements made on the same individual are likely to be more similar than measurements made on different individuals. Thus, observations on the same individual will not be independent. That is, repeated measurements on the same subjects are bound to be correlated (Diggle et al., 2002; Hox et al., 2010; Fitzmaurice et al., 2008; Yirga et al., 2020a). Thus, in this section, we review the general linear mixed model approach that can be extended for multivariate longitudinal data by assuming appropriate random effects. This method benefits from having extra correlation evolving from the longitudinal data structure that can be modeled within the same framework (Yirga et al., 2020a). Therefore, this study targeted identifying whether specific clinical and sociodemographic factors present in the data (and their respective possible interactions) influenced CD4 count in a cohort of HIV-infected patients receiving ART. The information and un-

derstanding of such elements are of epidemiological importance. The results will be beneficial in developing tools to support clinicians in identifying of factors related to HIV-infected patients (Yirga et al., 2020a). The results can be further use to shape communication and counseling strategies before treatment initiation.

2.2 Characteristics of longitudinal data

In longitudinal data, we have a high hierarchical structure. For example, consider the study of children's exam results within schools where the examination results of a random sample of students within a random sample of schools are compared. Here we have children within classrooms within schools (three levels) or just children within schools (two levels); patients within centers, and measurements within patients. We have got this high hierarchical structure. There is, of course, expected variation in our levels; children within one classroom will not all be the same. They will differ from one another on the outcome measure (exam score) or several other measures we could imagine. The schools will also be different; some will be in higher socioeconomic status in the neighborhood, and some will be at lower, or some schools will have better teachers than others. Because of these reasons, there is expected to see a variation in the outcome of the children (exam score of a child at the end of the child's school career). We will also expect some differences in the average exam score for the different schools.

Another possible example could be considering patients within nursing homes. We expect differences between patients and their outcomes. It is because one nursing home may be better at preventing a particular disease than another nursing home, or may be they have a different population of patients where we see different levels of a specific illness between centers. Measurements within patients will also differ from one another. But that daily or weekly, or monthly measurements will also show variation. The cluster of different 4 (CD4) cell count of a patient will change over time. If we measure HIV-infected patient's CD4 cell count today or next week, or next month, we will see variation. Therefore, measurements within patients also have variation at the measurement level and variation at the patient level. We also expect the "units" within a level will be correlated; for example children within a school will be more homogeneous and look more alike in terms of the same level of education and school activities than children from different schools. So we expect the exam scores to be somehow related to one another within a school. They have the same teachers, and they have the same type of education as the exam scores from children at a different school. The same can be said for longitudinal data. If we look at the CD4 cell count of a healthy person, the measurements will be very similar

over time; of course, we will see variation depends on many significant factors, but they will be very similar. Measurements of CD4 cell count within a person are more highly correlated than measurements from another CD4 cell count. We expect a correlation of the measurements when taking the "units" measurements within patients or children within classrooms; we expect these to be correlated (Liang et al., 2003).

Variables in multilevel data can be measured at different levels. At the second level (level-two), we could have the type of school: mixed school (boys and girls school) or single-gender school or university hospital or community hospital and so on. At the lower level, we can measure different variables: the reading ability of a child at intake, gender of the patient, age of the patient, etc. We can measure variables at both levels or two or three levels of data. It is essential to be able to use these variables measured at different levels in multilevel data analysis.

Multi-level data consist of multiple units of analysis, one nested within the other. There is a high hierarchical (multiple levels) data structure in multi-level data. Longitudinal or repeated measures data can be viewed as multi-level data, with repeated measures nested within individuals. In its simplest form, this leads to a two-level model, with the series of repeated measures at the lowest level and the individual subject at the highest level.

2.3 Mixed-Effects Models

Mixed-effects modeling is an advanced and vital method in statistics. It is a well-known method; therefore, we summarize the key aspects of the model relevant to the current study. The literature on mixed models is ubiquitous, and some of it can be found in here (Molenberghs & Verbeke, 2000; West et al., 2014; Littell et al., 2006; Searle et al., 2009; Pinheiro & Bates, 2006; Twisk, 2013; Liu, 2015; Hox et al., 2010; Rawlings et al., 2001; Hedeker & Gibbons, 2006; Duchateau et al., 1998; Fitzmaurice et al., 2008; Taris, 2000).

In longitudinal datasets, the response variables have more than one values per subject. This enables the analyst to study the improvement of the variable of interest within-subjects, in this manner eliminating the variation among subjects and thus increasing the power of the design. However, since observations on the same subject are almost always correlated, special techniques are required to deal with this dependence. Another way in which data can be dependent is when there is a hierarchical (multilevel) structure within the data, e.g., patients within the hospital, students within classrooms, etc. Mixed-effects models are one way of analyzing this

kind of data. This statistical method allows for the dependence of measurements in hierarchically structured data and independently examines the effects of variables at different levels. This study will discuss the use and theory of linear mixed-effects models that focus primarily on the continuous outcome variable.

2.3.1 Advantages of mixed-effects models

Using the mixed-effects model for longitudinal data helps to correctly account for the correlation of observations within a subject and quantify the heterogeneity between subjects due to unobserved factors. It is vital that before its implementation, the correct sample is determined based on prior information on the magnitude of the correlation. By correctly estimating the sample size, we end up with accurately estimated standard errors (SEs), which will give reliable confidence intervals (CI) and p-values (Yirga et al., 2020a). We can use the mixed-effects model to model variation at lower and higher levels of the design structure. Accounting for variation at a lower level gives us more power for estimation at a higher level (Hox et al., 2010; Yirga et al., 2020a). A mixed model is made up of fixed and random effects where the latter helps in accounting for correlation at a lower level within higher-level units. That is why mixed models are called "mixed" because the coefficients are a mix of fixed and random effects (Yirga et al., 2020a).

2.3.2 Fixed and random effects

In more general terms, fixed effects or fixed factors are covariates that we anticipate will influence the outcome variable. They are what we call explanatory variables in a standard linear regression. For instance, in our case, we are interested in making conclusions about how the socio-economic, demographic, and treatment type (place of residence, baseline BMI, baseline viral load, age, education level, marital status, HAART initiation, etc.) impacts the CD4 count of a patient. Therefore, these socio-economic, demographic, and treatment types are fixed effects, and the CD4 count of a patient is the response variable (Yirga et al., 2020a). Thus, a fixed-effect is the parameter of interest. The overall intercept is not the variable of interest, but of course, it is a fixed effect. In addition to the fixed effects, we also incorporate random effects in the mixed-effect model. Random effects are grouping factors for which we are attempting to control. A random intercept allows a different intercept for each subject. A random effect for a variable enables the effect of a variable on the outcome to differ between subjects. For example, a random effect could also be a random slope for a categorical variable. In general, in a mixed model, all of the variables of interest are added as fixed effects, but at least one and sometimes several of the fixed effects variables may also be added as random effects variables (Rawlings et al.,

2001; Yirga et al., 2020a). Therefore, the idea is that the random effects variables in the sample are a random sample of all possible variables in the broader population. Moreover, in longitudinal studies, time or a time-varying covariate X is often an explanatory variable of interest, and the associations between explanatory variables and responses may vary between subjects. A model that allows heterogeneity in the intercept and heterogeneity in the magnitude of the slope between subjects is referred to as the random intercept and slope model (Yirga et al., 2020a).

2.4 Linear mixed-effects model formulation

The random intercept and slope model is given by

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + b_{i0} + b_{i1} z_{ij} + \varepsilon_{ij}$$

where x_{ij} is the variable used as a predictor in the model.

A more general form of the mixed model is expressed as

$$y_{ij} = \beta_0 + \beta_1 x_{ij1} + \cdots + \beta_j p x_{ijp} + b_{i0} + b_{i1} z_{ij1} + \cdots + b_{iq} z_{ijq} + \varepsilon_{ij} \tag{2.1}$$

where y_{ij}, $j = 1, \cdots, n_i$, is the response of subject i at jth measurement, $\beta_i, \cdots, \beta_j p$ are the fixed effects coefficients, $x_{ij1}, \cdots, x_{ijp}$ are the fixed-effects regressors for subject i at jth measurement, $b_{i1}, \cdots, b_{iq}$ are the random-effects coefficients for subject i, $z_{ij1}, \cdots, z_{ijp}$ are the random-effects regressors, and ε_{ij} the error for subject i at jth measurement.

For our data analysis of CD4 cell count, we look at the square root of CD4 count as an outcome because it confirms that the model was better to the assumption of normally distributed residuals. Hence the model becomes

$$\sqrt{CD4_{ij}} = \beta_0 + \beta_1 x_{ij1} + \cdots + \beta_j p x_{ijp} + b_{i0} + b_{i1} z_{ijq} + \cdots + b_{iq} z_{ijq} + \varepsilon_{ij}$$

where $x_{i1}, \cdots, x_{ijp}$ are fixed effects.

The general matrix specification of the mixed model is

$$\underbrace{\boldsymbol{Y_i}}_{n_i \times 1} = \underbrace{\boldsymbol{X_i}}_{n_i \times p+1} \underbrace{\boldsymbol{\beta}}_{p \times 1} + \underbrace{\boldsymbol{Z_i}}_{n_i \times q+1} \underbrace{\boldsymbol{U_i}}_{q \times 1} + \underbrace{\boldsymbol{\varepsilon_i}}_{n_i \times 1} \tag{2.2}$$

with $i = 1, ..., n$ individuals and $j = 1, ..., n_i$ observations for individual i, where Y_i is $n_i \times 1$ vector of response variable, $X = [X_{ij1}, ..., X_{ijp}]$ is $n_i \times p + 1$ known design matrix that includes covariates for the fixed effects, β is $p \times 1$ vector of fixed effects parameters, Z_i is $n_i \times q + 1$ known design matrix (represents the observed values of covariates) for random effects (i^{th} subject), U_i is $q \times 1$ vector of random effects from a normal distribution with variance-covariance matrix G, and ε_i is $n_i \times 1$ error vector from a normal distribution with variance-covariance matrix R (Rawlings et al., 2001).

The assumption of the distribution of the random effects $U_i \overset{ind}{\sim} N(0, \sum_\nu)$, $\sum_\nu = \boldsymbol{G}$, and errors $\varepsilon_i \overset{ind}{\sim} N(0, \sigma^2 I_{n_i})$, $\sigma^2 I_{n_i} = \boldsymbol{R_i}$. Thus,

$$E \begin{bmatrix} \boldsymbol{U_i} \\ \boldsymbol{\varepsilon_i} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} and\ cov \begin{bmatrix} \boldsymbol{U_i} \\ \boldsymbol{\varepsilon_i} \end{bmatrix} = \begin{bmatrix} \boldsymbol{G} & 0 \\ 0 & \boldsymbol{R} \end{bmatrix} or \begin{bmatrix} \boldsymbol{U_i} \\ \boldsymbol{\varepsilon_i} \end{bmatrix} \sim N \left[\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{G} & 0 \\ 0 & \boldsymbol{R} \end{bmatrix} \right]$$

That is,

$$\boldsymbol{Y_i} \sim N(\boldsymbol{X_i\beta}, \boldsymbol{V} = \boldsymbol{ZGZ'} + \boldsymbol{R})$$

The variance-covariance matrix of $\boldsymbol{Y_i}$, $Var(\boldsymbol{Y_i}) = \boldsymbol{V}$, can be written as

$$\boldsymbol{V} = var(\boldsymbol{X_i\beta} + \boldsymbol{Z_iU_i} + \boldsymbol{\varepsilon_i}).$$

Since we assume that the random effects $\boldsymbol{U_i}$ and the errors $\boldsymbol{\varepsilon_i}$ are independent,

$$\boldsymbol{V} = var(\boldsymbol{X_i\beta}) + var(\boldsymbol{Z_iU_i}) + var(\boldsymbol{\varepsilon_i}).$$

Since β describes the fixed-effects parameters, $var(\boldsymbol{X_i\beta}) = 0$. Also, $\boldsymbol{Z_i}$ is a matrix of constant. Therefore,

$$\boldsymbol{V} = \boldsymbol{Z}var(\boldsymbol{U_i})\boldsymbol{Z'} + var(\boldsymbol{\varepsilon_i}).$$

Let $\boldsymbol{G}$ denote $var(\boldsymbol{U_i})$, and $var(\boldsymbol{\varepsilon_i}) = \boldsymbol{R}$. Hence,

$$\boldsymbol{V} = \boldsymbol{ZGZ'} + \boldsymbol{R}.$$

The general form of the $\boldsymbol{Z}$, $\boldsymbol{G}$, and $\boldsymbol{R}$ matrices can be shown as follows:

$$\boldsymbol{Z_i} = \begin{pmatrix} z_{i11} & z_{i12} & \cdots & z_{i1r} \\ z_{i21} & z_{i22} & \cdots & z_{i2r} \\ \vdots & \vdots & \ddots & \vdots \\ z_{in_i1} & z_{in_i2} & \cdots & z_{in_ir} \end{pmatrix}$$

$$Var(\boldsymbol{U_i}) = \boldsymbol{G} = \begin{pmatrix} Var(u_{1i}) & cov(u_{1i}, u_{2i}) & \cdots & cov(u_{1i}, u_{ri}) \\ cov(u_{1i}, u_{2i}) & Var(u_{2i}) & \cdots & cov(u_{2i}, u_{ri}) \\ \vdots & \vdots & \ddots & \vdots \\ cov(u_{1i}, u_{ri}) & cov(u_{2i}, u_{ri}) & \cdots & Var(u_{ri}) \end{pmatrix}$$

$$Var(\boldsymbol{\varepsilon_i}) = \boldsymbol{R_i} = \begin{pmatrix} Var(\varepsilon_{1i}) & cov(\varepsilon_{1i}, \varepsilon_{2i}) & \cdots & cov(\varepsilon_{1i}, \varepsilon_{n_i i}) \\ cov(\varepsilon_{1i}, \varepsilon_{2i}) & Var(\varepsilon_{2i}) & \cdots & cov(\varepsilon_{2i}, \varepsilon_{2i}) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\varepsilon_{1i}, \varepsilon_{n_i i}) & cov(\varepsilon_{2i}, \varepsilon_{n_i i}) & \cdots & Var(\varepsilon_{n_i i}) \end{pmatrix}$$

The distribution of Y is a multivariate normal distribution. The vector of n random variables $Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$ is said to have a multivariate normal distribution with mean vector $X\beta$ and variance-covariance non-singular matrix V. The probability density function (pdf) of the multivariate normal distribution is

$$f(\boldsymbol{Y}, \boldsymbol{\beta}, \boldsymbol{V}) = (2\pi)^{-n/2}|\boldsymbol{V}|^{-1/2} exp\left[\tfrac{-1}{2}(\boldsymbol{Y} - \boldsymbol{X\beta})'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X\beta})\right] \tag{2.3}$$

The log-likelihood of Y under this model is

$$\ell(\boldsymbol{\beta}, \boldsymbol{V}) = \frac{-n}{2} log(2\pi) - \frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X\beta})'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X\beta})$$

$$= \frac{-1}{2}\{n\log(2\pi) - \log|\boldsymbol{V}| + (\boldsymbol{Y} - \boldsymbol{X\beta})'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X\beta})\}$$

Therefore, the maximum likelihood estimate (MLE) of $(\boldsymbol{\beta}, \boldsymbol{V})$ is the one that maximizes the right-side of the above expression. To obtain the MLE of $\boldsymbol{\beta}$, for any fixed $\boldsymbol{V}$, differentiate the log-likelihood with regard to $\boldsymbol{\beta}$ both sides and equate to zero. Then replacing $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$ we solve for $\hat{\boldsymbol{\beta}}$:

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \frac{\partial}{\partial \beta}\left(\frac{-n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X\beta})'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X\beta})\right)$$

$$= \frac{\partial}{\partial \beta}\left(-\frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X\beta})'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X\beta})\right) = 0$$

$$\boldsymbol{X}'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X\beta}) = 0$$

$$\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Y} - \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X\beta} = 0$$

$$\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Y} = \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X\beta}$$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Y}$$

2.4.1 Covariance structure of repeated measures

The covariance structure of the observed data makes longitudinal data analysis particular (Searle et al., 2009; Yirga et al., 2020a). For the investigation to be substantial, the covariance among repeated measures must be demonstrated appropriately. The covariance structure is not the prime intrigued of study but is essential for significant inference. Covariance or correlation structures that are commonly used for longitudinal data analysis are compound symmetry (CS), unstructured (UN), First-order Autoregressive (AR(1)), and Toeplitz (Toep). These four common covariance structures are summarized here (Kincaid, 2005; Kowalchuk et al., 2004; Wolfinger, 1996; Kincaid, 2005; Hedeker & Gibbons, 2006; Rawlings et al., 2001; Searle et al., 2009; Little & Rubin, 2019; Hofer, 1998).

Compound Symmetry (CS)

Compound Symmetry correlation assumes observations of the same subject have homogeneous variance and homogeneous covariance. That is, both the variances and covariance across time are considered to be the same. For example, given four equally spaced time points, the CS has the following matrix structure:

$$CS = \begin{pmatrix} \sigma^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma^2 \end{pmatrix} = \sigma^2 \begin{pmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{pmatrix}$$

- Constant variance over time = $V(Y_{ij}) = \sigma^2$. This implies that all variances are assumed equal in CS.
- $C(Y_{ij}, Y'_{ij}) = \rho = \frac{\sigma_1^2}{\sigma^2}$, with σ_1^2 the variance within individuals, where $j \neq j'$. This implies that CS assumes an equal correlation between any two measurements of the same subject. ρ is then known as the intraclass correlation coefficient, a ratio of individual variance to the total variance.
- The CS or exchangeable correlation structure assumes correlations between all-time points to be equal, irrespective of the length of the time intervals.

Unstructured Correlation (UN)

All the variances and covariance in the UN are different over time (have no structure). This lets the data dictate what they should be and requires the estimate of

many parameters. But the more data that are used to assess the covariance structure, the fewer data are left to estimate the parameters of linear models. The UN matrix has the following form:

$$UN = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{pmatrix}$$

Unstructured correlation is a very flexible structure; however, this flexibility comes in with the price, and the price is having a lot of degrees of freedom.

Autoregressive of order 1 (AR (1)) correlation

In AR (1), observations per subject are assumed to be taken at equally-spaced intervals. The outcome has constant variance (σ^2) over time. AR (1) structure resolves some of the objectives to the use of CS. It uses a correlation between two responses that are m measurements apart is ρ^m; since ρ is $-1 \leq \rho \leq 1$, the greater the power, the smaller the magnitude. Thus, the further measurements are apart, the lower their correlation. The AR (1) structure is given by:

$$UN = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$$

TOEPLITZ (TOEP)

Similar to AR (1) in that all the correlations at the same distance have the same relationship; but no assumption of exponential decay. The AR (1) model can be estimated with a single parameter (and then the exponent of the distance). The Toeplitz model has many settings due to distance.

$$TOEP = \begin{pmatrix} \sigma^2 & \sigma^2_{12} & \sigma^2_{13} & \sigma^2_{14} \\ \sigma^2_{21} & \sigma^2 & \sigma^2_{21} & \sigma^2_{22} \\ \sigma^2_{31} & \sigma^2_{32} & \sigma^2 & \sigma^2_{31} \\ \sigma^2_{41} & \sigma^2_{42} & \sigma^2_{43} & \sigma^2 \end{pmatrix}$$

2.4.2 Spatial covariance structure in mixed-effects models

Spatial covariance structure measures the actual distance or variation among observations in space that are identified as unequally spaced longitudinal data (Zimmerman & Harville, 1991; Littell et al., 2006). The objective of including spatial covariance structure in mixed-effects models is to account for spatial variability (heterogeneity), failure to do so can result in erroneous conclusions. The spatial covariance structure model, written as

$$C(h) = C_0 + \sigma^2 \rho(h) \tag{2.4}$$

where C_0, σ^2 and $\rho(h)$ indicates the *nugget*, the *sill* and the *range* (covariance structure model), respectively (Zimmerman & Harville, 1991; Littell et al., 2006).

The four commonly used covariance structure for longitudinal data analysis described in Subsection (2.4.1) is used when the time points are equally spaced. In many longitudinal data, repeated measurements are not designed to have equal intervals, or some subjects may enter a follow-up survey after a specified interview data due to sickness, migration, or some other reasons (Liu, 2015). In these circumstances, the use of covariance pattern models discussed in Subsection (2.4.1) is no longer reasonable; instead, spatial covariance structures that take into account the distance between the observations within each subject can be used. There are four commonly used spatial covariance structures to analyze longitudinal data with unequal time intervals: spatial exponential, spatial power, spatial spherical, and spatial Gaussian. Each of these spatial covariance pattern models is based on the assumption that correlations between measurements are positive and decreasing functions of the Euclidean distance, $\hat{d}_{ij}$, which is defined as the absolute difference between two-time points, where $i \neq j$ and $i > j \quad (i, j = 1, ..., n)$ (Liu, 2015). For illustrative convenience and simplicity, spatial covariance pattern models for four unequally spaced time points are presented.

The spatial power covariance structure, SP(POW), is given by

$$\boldsymbol{R} = \sigma^2 \begin{pmatrix} 1 & \rho^{d_{12}} & \rho^{d_{13}} & \rho^{d_{14}} \\ \rho^{d_{21}} & 1 & \rho^{d_{23}} & \rho^{d_{24}} \\ \rho^{d_{31}} & \rho^{d_{32}} & 1 & \rho^{d_{34}} \\ \rho^{d_{41}} & \rho^{d_{42}} & \rho^{d_{43}} & 1 \end{pmatrix}$$

The spatial exponential covariance pattern model (SP(EXP)), which is an extension of the spatial power covariance structure, with $d_{ij} = d_{ji}$, is given by

$$\boldsymbol{R} = \sigma^2 \begin{pmatrix} 1 & exp(-d_{12}/\rho) & exp(-d_{13}/\rho) & exp(-d_{14}/\rho) \\ exp(-d_{21}/\rho) & 1 & exp(-d_{23}/\rho) & exp(-d_{24}/\rho) \\ exp(-d_{31}/\rho) & exp(-d_{32}/\rho) & 1 & exp(-d_{34}/\rho) \\ exp(-d_{41}/\rho) & exp(-d_{42}/\rho) & exp(-d_{43}/\rho) & 1 \end{pmatrix}$$

The spatial Gaussian covariance structure, or SP(GAU), has the following structure:

$$\boldsymbol{R} = \sigma^2 \begin{pmatrix} 1 & exp(-d_{12}^2/\rho^2) & exp(-d_{13}^2/\rho^2) & exp(-d_{14}^2/\rho^2) \\ exp(-d_{21}^2/\rho^2) & 1 & exp(-d_{23}^2/\rho^2) & exp(-d_{24}^2/\rho^2) \\ exp(-d_{31}^2/\rho^2) & exp(-d_{32}^2/\rho^2) & 1 & exp(-d_{34}^2/\rho^2) \\ exp(-d_{41}^2/\rho^2) & exp(-d_{42}^2/\rho^2) & exp(-d_{43}^2/\rho^2) & 1 \end{pmatrix}$$

The variance-covariance structure for the spatial spherical pattern model (SP(SPH)) is given by

$$= [1 - 1.5(d_{ij}/\rho) + 0.5(d_{ij}/\rho)^3] \times 1\{d_{ij} < \rho\},$$

where the function $1\{d_{ij} < \rho\}$ is an indicator function that equal 1 when $d_{ij} < \rho$ and equals 0 otherwise (Littell et al., 2006).

2.4.3 Semivariogram

A good measure of the spatial continuity of **z(s)** is defined by means of the variance of the difference $\boldsymbol{z(t_i) - z(t_j)}$, where t_i and t_j are unequally spaced time points in d in the context of longitudinal data. Specifically, consider t_i and t_j to be spatial increments such that $h = t_j - t_i$, then the variance function based on the increments h is independent of the time points, t_i, t_j. According to Cressie (2015), most commonly, the continuity measure used in practice is one half of this variance, also known as the semivariogram (semivariance) function,

$$\begin{aligned} \boldsymbol{\gamma_z(h)} &= \frac{1}{2} Var\left[z(t+h) - z(t)\right] \\ &= \frac{1}{2}\left(E\{[z(t+h) - z(t)]^2\} - \{E[z(t+h)] - E[z(t)]\}^2\right) \end{aligned} \tag{2.5}$$

The semivariogram, based on either the random intercept or the random coefficient model, is usually applied for spatial data when the time intervals are unequally spaced (Liu, 2015). Since the empirical semivariogram is sensitive to outliers, influence diagnostics need to be performed first before fitting a smooth curve to the scatter-plot of the empirical semivariogram.

The approximate empirical semivariance with a Gaussian-type form is

$$\boldsymbol{\gamma_z(h)} = C_0 \left[1 - exp\left(-\frac{h^2}{a_0^2}\right)\right] \tag{2.6}$$

where $\boldsymbol{\gamma_z(h)}$ is the semivariance function, $C_0 = C_n + \sigma_0^2$ is the *sill* consists of the *nugget* effect (C_n), if present, and the partial sill σ_0^2 (Cressie, 2015).

The commonly used theoretical semivariogram shape rises monotonically as a function of distance. The shape is typically characterized in terms of particular parameters; these are the range (a_0), the sill (or scale, C_0), and the nugget effect.

Specifically, the *sill* is the semivariogram upper bound. The *range* represents the distance at which the semivariogram reaches the sill. When the semivariogram increases asymptotically towards its sill value, as occurs in the exponential and Gaussian semivariogram models, the term effective (practical) range is used. The effective range r_ϵ is defined as the distance at which the semivariance value achieves 95% of the sill. In particular, for these models the relationship between the range and effective range is $r_\epsilon = 3a_0$ (exponential model) and $r_\epsilon = \sqrt{3a_0}$ (Gaussian model) (Littell et al., 2006). The nugget effect C_n represents a discontinuity of the semivariogram that can be present at the origin. It is typically attributed to microscale effects or measurement errors. The semivariance is always 0 at distance $\mathbf{h} = 0$; hence, the nugget effect demonstrates itself as a jump in the semivariance as soon as $\mathbf{h} > 0$ (SAS, 2014).

The four commonly used spatial covariance pattern models in terms of the semivariogram: power, exponential, spherical, and Gaussian are given here

Power:

$$\boldsymbol{\gamma_z(h)} = \begin{cases} 0 & \text{if} \quad |h| = 0 \\ C_n + \sigma_0^2 h^{a_0} & \text{if} \quad 0 < |h|, \quad 0 \leq a_0 < 2 \end{cases}$$

Exponential:

$$\boldsymbol{\gamma_z(h)} = \begin{cases} 0 & \text{if} \quad |h| = 0 \\ C_n + \sigma_0^2 \left[1 - exp\left(-\frac{|h|}{a_0}\right)\right] & \text{if} \quad 0 < |h| \end{cases}$$

Spherical:

$$\gamma_{\boldsymbol{z}}(\boldsymbol{h}) = \begin{cases} 0 & \text{if} \quad |h| = 0 \\ C_n + \sigma_0^2 \left[1.5|h|/a_0 - 0.5\left(|h|/a_0\right)^3\right] & \text{if} \quad 0 < |h| \leq a_0 \\ C_0 & \text{if} \quad a_0 < |h| \end{cases}$$

Gaussian:

$$\gamma_{\boldsymbol{z}}(\boldsymbol{h}) = \begin{cases} 0 & \text{if} \quad |h| = 0 \\ C_n + \sigma_0^2 \left[1 - exp\left(-\frac{|h|^2}{a_0^2}\right)\right] & \text{if} \quad 0 < |h| \end{cases}$$

A detailed discussion of various spatial covariance pattern models in terms of the semivariogram can be found in the literature by Cressie (2015), and SAS (2014).

2.4.4 Model selection in mixed models

To decide which mixed-effects model are fits the data best, we can use likelihood-based methods, i.e., either the likelihood ratio test (LRT) or Information Criteria (IC) such as Akaike Information Criteria (AIC) or Bayesian Information Criteria (BIC) method. The LRT, which is based on distribution, can be used to test nested models. The model with the smallest AIC and BIC (the one with the highest likelihood given the parameters in the model) is the best fitting model. That is, the AIC and BIC can be used to compare models such that the smaller of any of these, the better between two or more competing models. The IC method is more general to compare two or more competing non-nested models. However, the LRT is the best method to compare nested models (Loy et al., 2017).

Regarding the variance-covariance structure, we have many choices. We have discussed the four most appropriate correlation structures for longitudinal data analysis in the previous section. Ideally, the covariance structure should be known from previous worth or subject matter consideration. Otherwise, one should look for a structure that gives a better fit. We contemplate a few likely structures and choose among them according to some measures fit. These measures have two components: one that rewards for accuracy of fit and another that penalizes for the number of parameters it takes to achieve it. The most popular of these methods are arranged in tabular form below.

Table 2.1: Summary of Information Criteria

Criteria	Large is better	Small is better	Reference
AIC	ℓ-d	-2ℓ+2d	Akaike (1974)
BIC	$\ell - \frac{1}{2}$d log n	-2ℓ+d log n	Schwarz et al. (1978)
CAIC	$\ell - \frac{d(logn+1)}{2}$	-2ℓ+d(log n+1)	Bozdogan (1987)
HQIC	ℓ-d log log n	-2ℓ+2d log log n	Hannan & Quinn (1979)

2.4.5 Parameter estimation in mixed models

Let Y_i denote the vector of observations from one individual. $Y_i = (y_{i1}, y_{i2}, \cdots, y_{in_i})$ assuming n_i observations per individual. Variance-covariance matrix for this is V_i and mean is $X_i\beta$ where X_i is the design matrix and β is the vector of observations.

In mixed models, we use maximum likelihood (ML) to estimate the fixed effects, the standard errors of the fixed effects, and the variance of the random effects. The likelihood of mixed effect models can be time-consuming computationally, but with advances in statistical software, this has become an easily manageable problem. Often the likelihood is solved by iteration until convergence. However, under ML estimation, the residual variance and variance of random effects are underestimated. Thus, instead, the restricted maximum likelihood (REML) estimation gives unbiased estimates of variance parameters by taking into account the degrees of freedom (DF) utilized to estimate the fixed effects; hence variance parameter estimates are generally larger than those from ML estimation. However, REML uses the covariate mean structure (the number of fixed effects) in the model to adjust. That means we use REML when we are comparing two models that differ only in random effects (Rawlings et al., 2001; Littell et al., 2006; Yirga et al., 2020a).

In general, when testing mixed-effects models that differ in variance components, we could either use REML or ML since they both give interpretable LRT and IC for such a comparison. However, testing and comparing models that differ in fixed effects, then only ML, will provide us with interpretable LRT and IC. However, ML does not take into account the degrees of freedom for the loss of fit in the estimation of parameters, but REML does (Rawlings et al., 2001; Hofer, 1998). REML is the default of SAS PROC MIXED and PROC GLIMMIX for mixed models with normally distributed data; the details are given in Littell et al. (2006).

Maximum Likelihood (ML) Estimation

ML estimation is a method of obtaining estimates of model parameters by minimizing the likelihood function. The likelihood function, L, measures the likelihood of unknown parameters given the observations and is defined using the density function of the data (Brown & Prescott, 2014; Hedeker & Gibbons, 2006). In statistical models where the data are assumed to be independent (e.g., fixed-effects models), L is simply the product of each observation's density functions. However, in a mixed-effects model, observations are not independent, and L needs to be based on a multivariate normal density function (Equation 2.3) for the data (West et al., 2014; Hedeker & Gibbons, 2006). The Corresponding log-likelihood function, ℓ, is also described in Equation 2.3.

Parameters in a specified model are fixed unknown constants to be estimated from the data. The parameters in Equation (2.3) are thus vector of fixed effects $\boldsymbol{\beta}$, and all unknowns in the variance-covariance matrices $\mathbf{G}$ and $\mathbf{R}$. All unknowns in the $\mathbf{G}$ and $\mathbf{R}$ matrices are collectively referred to as the *covariance parameters* and denoted as $\boldsymbol{\theta}$ (Littell et al., 2006).

In ML estimation, although it is possible to find estimates of $\boldsymbol{\beta}$, and $\boldsymbol{\theta}$ simultaneously, by maximizing the log-likelihood of Equation (2.3) with respect to both $\boldsymbol{\beta}$, and $\boldsymbol{\theta}$, many iterative algorithms simplify the maximization by profiling-out the $\boldsymbol{\beta}$ parameters from ℓ Equation (2.3). The most common iterative methods used for the maximization problem in the context of mixed-effects models are Expectation-Maximization (EM), Newton-Raphson (N-R), and Fisher scoring method. For detailed reviews of these methods, see West et al. (2014), Hedeker & Gibbons (2006), Casella & Berger (2002), Harville & Callanan (1990), Harville (1977) or Searle et al. (2009).

ML Estimation for known $\boldsymbol{\theta}$

Consider a special case of ML estimation for linear mixed-effects models that all parameters in $\mathbf{G}$ and $\mathbf{R}$, and the matrix $\mathbf{V} = \mathbf{R} + \mathbf{ZGZ'}$ are assumed to be known.

Since $\boldsymbol{\theta}$ is assumed to be known, the only parameters that we estimate are the fixed-effects, $\boldsymbol{\beta}$. Therefore, optimization of $\boldsymbol{\beta}$ is equivalent to finding a minimum of an objective function $\frac{\partial \ell}{\partial \boldsymbol{\beta}}$, defined by the last term in Equation (2.3) and setting the re-

sulting expression to zero:

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \left(\frac{-1}{2}(Y - X\beta)' V^{-1}(Y - X\beta) \right) = 0$$

Note that optimization of $\frac{\partial \ell}{\partial \boldsymbol{\beta}}$ with respect to $\boldsymbol{\beta}$ can be carried out by applying the method of generalized least square (GLS) (West et al., 2014). Rearrangement of the above equation gives the ML estimate of the parameter $\boldsymbol{\beta}$ for a known $\boldsymbol{\theta}$:

$$\hat{\boldsymbol{\beta}} = (X' V^{-1} X)^{-1} X' V^{-1} Y$$

where the estimate $\hat{\boldsymbol{\beta}}$ has the desirable statistical property of being the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$. For a detailed description of this property can be found in West et al. (2014), Brown & Prescott (2014), McCulloch et al. (2008), and Christensen (1991).

ML Estimation for unknown $\boldsymbol{\theta}$

With the assumption of the covariance parameter, $\boldsymbol{\theta}$, unknown but not being a function of the fixed effects, $\boldsymbol{\beta}$, the log-likelihood function of Equation (2.3) has to be maximized with respect to $\boldsymbol{V}$ (unknown parameters describing $\mathbf{G}$ and $\mathbf{R}$). The ML equation for $\boldsymbol{V}$ can be obtained by taking partial derivatives of Equation (2.3) with respect to $\boldsymbol{V}$ and setting the resulting equation equal to Zero, using $\boldsymbol{\theta}$ for each parameter in $\boldsymbol{V}$. But $\boldsymbol{\beta}$ is implicitly a function of $\boldsymbol{V}$.

The log-likelihood for $\boldsymbol{\beta}$, and $\boldsymbol{\theta}$ is written as

$$\ell_{(\boldsymbol{\beta},\boldsymbol{\theta};\boldsymbol{Y})} = \frac{-n}{2} log(2\pi) - \frac{1}{2} log|\boldsymbol{V}(\boldsymbol{\theta})| - \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})' \boldsymbol{V}(\boldsymbol{\theta})^{-1} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{2.7}$$

The partial derivatives can be solved by using some results of matrix and vector differentiation. Christensen (1991) and McCulloch et al. (2008) presented the following four results:

1. $\frac{\partial Ax}{\partial x} = A$
2. $\frac{\partial x' Ax}{\partial x} = 2x' A$
3. If A is a function of a scalar s,
$$\frac{\partial A^{-1}}{\partial s} = -\frac{A^{-1} \partial A}{\partial s} A^{-1}$$
4. If A is a function of a scalar s,
$$\frac{\partial log|A|}{\partial s} = tr\left(A^{-1} \frac{\partial A}{\partial s}\right).$$

Additional determinants and inverse properties that are useful for the differentiation of matrix expression are highlighted in Appendix C. The following partial derivatives are obtained using the above four matrix and vector differentiation results:

$$\begin{aligned}\frac{\partial \ell}{\partial \boldsymbol{\beta}} &= -\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{V}(\boldsymbol{\theta})^{-1}\boldsymbol{X} + \boldsymbol{Y}'\boldsymbol{V}(\boldsymbol{\theta})^{-1}\boldsymbol{X}, \qquad \text{and} \\ \frac{\partial \ell}{\partial \theta_j} &= -\frac{1}{2}\left[tr\left(\boldsymbol{V}^{-1}\frac{\partial V}{\partial \theta_j}\right) - (\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}(\boldsymbol{\theta})^{-1}\frac{\partial V}{\partial \theta_j}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})\right],\end{aligned} \tag{2.8}$$

where $j = 1, ..., q$. The above partial derivatives are set equal to zero to get the following set of estimating equation which can be solved to obtain the ML estimates of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$:

$$\begin{aligned}\boldsymbol{X}'\boldsymbol{V}(\boldsymbol{\theta})^{-1}\boldsymbol{X}\boldsymbol{\beta} &= \boldsymbol{X}'\boldsymbol{V}(\boldsymbol{\theta})^{-1}\boldsymbol{Y} \\ tr\left(\boldsymbol{V}^{-1}\frac{\partial V}{\partial \theta_j}\right) &= (\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}(\boldsymbol{\theta})^{-1}\frac{\partial V}{\partial \theta_j}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})\end{aligned} \tag{2.9}$$

Since the above estimating equation does not have simple closed-form solutions, they can be solved simultaneously by using iterative methods to obtain ML estimates of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$. The conventional optimization methods which require first and second derivatives (e.g., Newton-Raphson and Fisher scoring) may be applied. However, instead of solving Equation (2.9) simultaneously, an alternative method of maximizing Equation (2.7), which is often more convenient, is the method of *profile-likelihood*. Thus, one can evaluate the profile-likelihood for $\boldsymbol{V}$ denoted ℓ_P, which is the likelihood for a given value of $\boldsymbol{V}$ with the maximizing value of $\boldsymbol{\beta}$ for that $\boldsymbol{V}$ is inserted.

$$\ell_P = -\frac{1}{2}\boldsymbol{Y}P_{\boldsymbol{Y}} - \frac{1}{2}\log|\boldsymbol{V}| - \frac{n}{2}\log(2\pi), \tag{2.10}$$

where $P = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}$. For the presentation of efficient computational methods for maximizing ℓ_P, see Searle et al. (2009), Searle (1982), and Pinheiro & Bates (2006).

An advantage of ML estimators is their efficiency-they simultaneously utilize all of the available data and account for any dependence. The limitation with a variance-component estimation through the usual ML approach is that all fixed-effects are assumed to be known without error. This is not always true in practice, and as a consequence, ML estimators yield biased estimates of variance components. Most notably, estimates of the residual variance tend to be underestimated. This bias occurs because the ML estimates of $\boldsymbol{\theta}$ do not take into account the loss of a degree of freedom (information used up) that results from estimating the fixed-effect parameters in $\boldsymbol{\beta}$ (West et al., 2014; Rawlings et al., 2001; Hedeker & Gibbons, 2006; Hofer, 1998). This can be illustrated with a simple scenario. Consider a simple random

sample, $x_1, ..., x_n$, identically and independently distributed $N(\mu, \sigma^2)$, then the ML variance estimator of σ^2 is $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ rather than the unbiased Analysis of Variance (ANOVA) estimator $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. In estimating the variance, an ML estimator ignores the fact that parameters in the mean have been estimated. One degree of freedom is used up in estimating the population means with $\bar{x}$; therefore, the appropriate divisor is supposed to be $n - 1$ (the number of observations minus number of *non-redundant* parameters estimated) rather than n.

The bias in ML estimates can also become quite large when a model contains many fixed effects than the sample size (Searle et al., 2009). A detailed discussion of the bias in ML estimates of $\boldsymbol{\theta}$ in the context of mixed-effects models is provided by Molenberghs & Verbeke (2000). An alternative way of the ML method known as Residual (restricted) maximum likelihood (REML) estimation, which was first suggested by Patterson & Thompson (1971), is frequently used to eliminate the bias in ML estimates of $\boldsymbol{\theta}$. We discuss REML estimation in the following subsection.

Restricted Maximum Likelihood (REML) Estimation

REML estimation is an alternative way of estimating the covariance parameters in $\boldsymbol{\theta}$. REML is often preferred to ML estimation since it produces unbiased estimates of covariance parameters by taking into account the degrees of freedom lost, which results from estimating the fixed effects $\boldsymbol{\beta}$ (West et al., 2014; Rawlings et al., 2001). REML aims to improve upon the ML estimator of $\boldsymbol{\theta}$, not all of the model's parameters. However, given a REML estimator of $\boldsymbol{\theta}$, it is evident how the ML estimator of $\boldsymbol{\beta}$ should be formed.

Unlike ML estimators, REML estimators maximize only the part of the likelihood, which is invariant to $\boldsymbol{X\beta}$. In this sense, REML is a restricted version of ML. That is, $\boldsymbol{\beta}$ is eliminated from the log-likelihood by considering the likelihood of a set of the linearly transformed response data vector, whose distribution does not contain any fixed effects, rather than the density of the response vector (Y) itself. Harville (1977) refers to linearly transformed $K'Y$ for K' of this nature as being *error contrast*, where K' is any $(n-p) \times n$ matrix of full rank satisfying $K'X = 0$; $E(K'Y) = 0 \quad \forall \beta$. Detail discussions of error contrast have appeared in Patterson & Thompson (1971), Harville (1977), Searle et al. (2009), and Searle (1982).

REML Equations

Searle et al. (2009) shows the REML estimation of $\boldsymbol{\theta}$, which is summarized as fol-

lows:
Recall the linear mixed model:

$$Y = X\beta + ZU + \varepsilon,$$

where $Y \sim N(X\beta, V)$. Consider the set of values $K'Y$ where vectors of the form K' can be chosen to satisfy

$$K'Y = K'X\beta + K'ZU \tag{2.11}$$

such that no term in β is contained. That is $K'X\beta = 0 \quad \forall\beta$. Therefore, $K'Y \sim N(0, K'VK)$. Then,

$$\begin{aligned} K'Y &= K'ZU \\ E(K'Y) &= E(K'ZU) = K'E(ZU) = K'ZE(U) = 0. \end{aligned}$$

This confirms that REML estimation has no procedure for fixed effects.

The information matrix is essential to REML in that it plays a role in estimating the variance components. Searle et al. (2009) also shows the information matrix, which is also summarized as follows:

$$\begin{aligned} \ell_Y &= -\frac{n}{2}log(2\pi) - \frac{1}{2}log|V| - \frac{1}{2}(Y - X\beta)'V^{-1}(Y - X\beta) \\ \ell_{K'Y} &= -\frac{n-r}{2}log(2\pi) - \frac{1}{2}log|K'VK| - \frac{1}{2}(K'Y - 0)'|K'VK|^{-1}(K'Y - 0) \\ &= -\frac{n-r}{2}log(2\pi) - \frac{1}{2}log|K'VK| - \frac{1}{2}(K'Y)'|K'VK|^{-1}(K'Y) \\ &= -\frac{n-r}{2}log(2\pi) - \frac{1}{2}log|K'K| - \frac{n-r}{2}log|V| - \frac{(Y'K)}{2|V|}|K'K|^{-1}(K'Y) \end{aligned} \tag{2.12}$$

$$\begin{aligned} \frac{\partial \ell_{K'Y}}{\partial V} &= -\frac{n-r}{2V} + \frac{1}{2V^2}(Y'K')(K'K)^{-1}(K'Y) = 0 \\ \frac{n-r}{2V} &= \frac{1}{2V^2}(Y'K')(K'K)^{-1}(K'Y) \\ (n-r)V &= (Y'K')(K'K)^{-1}(KY) \\ \hat{V} &= \frac{(Y'K')(K'K)^{-1}(KY)}{n-r}, \end{aligned} \tag{2.13}$$

where $K'(K'K)^{-1}K = I_n - X(X'X)^{-1}X'$ (Searle et al., 2009). Therefore, Equation (2.13) can be expressed as:

$$\hat{V} = \frac{Y'(I_n - X(X'X)^{-1}X')Y}{n-r}.$$

Let $X'X = n$, which results $(X'X)^{-1} = \frac{1}{n} = n^{-1}$, and $XX' - I_n$. Substitute these in the above expression, results:

$$\hat{V} = \frac{Y'(I_n - n^{-1}I_n)Y}{n-r}, \qquad \text{which is the REML estimate of} \quad V.$$

Recall: $Var(K'Y) = Var(K'ZU)$
By definition: $Var(Y) = E((Y - E(Y))(Y - E(Y))')$
Therefore,

$$\begin{aligned} Var(K'Y) &= E((K'Y - E(K'Y))(K'Y - E(K'Y))') \\ &= K' \underbrace{E((Y - E(Y))(Y - E(Y))')}_{\text{V}} K \\ &= K'VK, \qquad \text{which is the REML variance component.} \end{aligned} \tag{2.14}$$

Refer to Searle et al. (2009) for explicit formula of REML estimating equations.

The advantage of having the likelihood and IC in mixed-effects models is that we can easily compare different models, we can discern what the effect is adding or removing a fixed effect (ML approach) or random effect (either ML or REML approach) by comparing the AIC, BIC and the likelihoods, where for the AIC and BIC a lower value is better, means that the model fits better. We could also compare models using $-2LogLikelihood$ and the $Likelihood\ ratio\ test$. Then, if a model is significantly lower, it means that it is better if models are not significantly different from each other, then we prefer the simpler of the two models. A model with fewer parameters is not significantly worse than a more complex model, is considered to be better. Note that a simpler model cannot be significantly better; it is just that it is not significantly worse than a complex model. It is more straightforward, and that is why we prefer if models are not significantly different from each other.

ML and REML are the two most common methods available to estimate the parameters ($\beta_i's$ and $b_i's$) in mixed models, as discussed in the above section. Our preference is for the REML approach because it takes into account the DF in the VC estimation, and it gives us unbiased estimates. Several software packages make it possible to perform mixed-effects models in R, such as the MASS package with the function glm (generalized linear model), lme (linear mixed-effects) function in the $nlme$ (nonlinear mixed-effects) package, and $lmer$ (linear mixed-effects regression) function in the $lme4$ (linear mixed-effects with S4 classes) package. We could also use SAS Proc Mixed, Stata, and SPSS to do a mixed model.

2.5 Residual and Influence Diagnostics

Under linear mixed-effects models, the distributional assumptions for the random effects $\boldsymbol{b_i}$, and the residuals $\boldsymbol{\varepsilon_i}$, are assumed to be satisfied. However, this may not be true in practice because, in the parameter estimation of linear mixed-effects models, some bizarre observations can have undue influences on the chosen model. Therefore, once a chosen mixed-model is fitted with longitudinal data, it is necessary to carry out model diagnostics for verifying whether distributional assumptions for the residuals are satisfied or meet various assumptions on model specifications (Liu, 2015; West et al., 2014; Schabenberger, 2005).

In longitudinal data analysis, plotting various residuals reveals inconsistencies between the observed data and the model-based predictions (Diggle et al., 2002). Moreover, the identification of influential observations also needs to be performed in longitudinal data analysis to check whether the fit of the model is sensitive to unusual observations (Liu, 2015; Diggle et al., 2002; West et al., 2014; Schabenberger, 2005). This notion of a standard diagnostic technique is discussed in this section. We focus on the definitions of a selected set of terms related to residual and influence diagnostics in the context of linear mixed-effects models. A detailed discussion of existing diagnostic techniques can be found in numerous literature (Cook & Weisberg, 1982; Liu, 2015; Diggle et al., 2002; West et al., 2014; Schabenberger, 2005).

2.5.1 Residual Diagnostics

Recall the general linear mixed model

$$\boldsymbol{y_i} = \boldsymbol{x_i'\beta} + \boldsymbol{z_i'u_i} + \boldsymbol{\varepsilon_i}.$$

Let the predicted mean response for subject i at time point j be $\boldsymbol{\hat{\mu}_{ij}} = \boldsymbol{x_i'\hat{\beta}}$. Then, a residual is the deviation of the predicted value from the observed value. In the mixed model, residuals are distinguished as *marginal* and *conditional*. An n_i-dimensional vector of residuals for each subject can then be obtained from an LMM, given by

$$\boldsymbol{r}_{mi} = \boldsymbol{y_i} - \boldsymbol{x_i'\hat{\beta}},$$

where $\boldsymbol{r}_{mi}$ indicates the marginal residuals. If the random-effects model is considered, residuals for each subject become conditional on the random effects, written as

$$\boldsymbol{r}_{ci} = \boldsymbol{y_i} - \boldsymbol{x_i'\hat{\beta}} - \boldsymbol{z_i'u_i} = \boldsymbol{r}_{mi} - \boldsymbol{z_i'u_i},$$

where $\boldsymbol{r}_{ci}$ indicates the conditional residuals Liu (2015).

Residuals are used to verify model assumptions, detect outliers, and identify potentially influential observations. In general, residuals are useful for assessing normality, constant variance, and outliers. However, the raw residuals ($\boldsymbol{r}_{mi}$ and $\boldsymbol{r}_{ci}$) in the context of LMMs, are less suited for these purposes because the raw residuals will exhibit correlations and have heterogeneous variances. Therefore, studentized and Pearson residuals can be considered since they account for the unequal variance of the residuals (West et al., 2014; Schabenberger, 2005; Littell et al., 2006). Adjusting the raw residuals by their actual standard deviations obtain standardized residuals. However, the actual standard deviations are rarely known in practice; therefore, adjusting is done using estimated residual variances, which are then referred to as studentized residuals. Raw residuals can also be adjusted by their estimated variances of the observed response $\boldsymbol{y_i}$, referred to as Pearson residuals (West et al., 2014; Liu, 2015).

Given the specifications of linear mixed model, Liu (2015) noted that the variance-covariance matrix of $\boldsymbol{r}_{mi}$ is

$$Var(\hat{\boldsymbol{r}}_{mi}) = \hat{\boldsymbol{V}}_{\boldsymbol{i}} - \boldsymbol{x}_i(\boldsymbol{x}_i'\hat{\boldsymbol{V}}_i^{-1}\boldsymbol{x}_i)^{-1},$$

where $\hat{\boldsymbol{r}}_{mi}$ is the estimated total variance of $\boldsymbol{y}_i$, from either MLE or the REML estimator. The variance-covariance matrix of conditional residuals can be specified according to Grégoire et al. (1995) as follows:

$$Var(\hat{\boldsymbol{r}}_{ci}) = (\boldsymbol{I}_{n_i} - \boldsymbol{Z}_i\hat{\boldsymbol{G}}\boldsymbol{Z}_i'\hat{\boldsymbol{V}}_i^{-1})Var(\hat{\boldsymbol{r}}_{mi})(\boldsymbol{I}_{n_i} - \boldsymbol{Z}_i\hat{\boldsymbol{G}}\boldsymbol{Z}_i'\hat{\boldsymbol{V}}_i^{-1})',$$

Another method of adjusting residuals, rather than dividing each individual residual by the variance of an observation, is to consider the vector of residuals and the estimated variance $\boldsymbol{V}(\hat{\boldsymbol{\theta}})$. Let $\hat{\boldsymbol{C}}$ denote a matrix such that $\hat{\boldsymbol{C}}\hat{\boldsymbol{C}}' = \boldsymbol{V}(\hat{\boldsymbol{\theta}})$, its lower-triangular Cholesky root (see Appendix C). Then the adjusted residuals $\boldsymbol{r}_c = \hat{\boldsymbol{C}}^{-1}\boldsymbol{r}_m$ have zero mean and are approximately uncorrelated (West et al., 2014). Table 2.2 shows summary of residuals.

2.5.2 Influence Diagnostics

In determining parameter estimates and other statistics, it is well known that not all observations in a data set play an equal role (Zewotir & Galpin, 2005). Identification of influential observations is essential to identify particular observations that have extraordinary influences on the analytic results (West et al., 2014; Liu, 2015). Specifi-

Table 2.2: Summary of residuals in LMMs

Type of Residuals	Marginal	Conditional
Raw	$\boldsymbol{r}_{mi} = \boldsymbol{y_i} - \boldsymbol{x}_{\boldsymbol{i}}'\hat{\boldsymbol{\beta}}$	$\boldsymbol{r}_{ci} = \boldsymbol{r}_{mi} - \boldsymbol{z}_{\boldsymbol{i}}'\hat{\boldsymbol{u}}_{\boldsymbol{i}}$
Studentized	$\boldsymbol{r}_{mi}^S = \frac{\boldsymbol{r}_{mi}}{\sqrt{\hat{Var}(\boldsymbol{r}_{mi})}}$	$\boldsymbol{r}_{ci}^S = \frac{\boldsymbol{r}_{ci}}{\sqrt{\hat{Var}(\boldsymbol{r}_{ci})}}$
Pearson	$\boldsymbol{r}_{mi}^P = \frac{\boldsymbol{r}_{mi}}{\sqrt{\hat{Var}(\boldsymbol{y}_i)}}$	$\boldsymbol{r}_{ci}^P = \frac{\boldsymbol{r}_{ci}}{\sqrt{\hat{Var}(\boldsymbol{y}_i\|\boldsymbol{u}_i)}}$
Scaled	$\boldsymbol{R}_{mi} = \hat{C}^{-1}\boldsymbol{r}_{mi}$	

cally, influence diagnostics on longitudinal data involve individuals having multiple data points rather than at a single time. Consequently, the removal of one individual can affect a series of observations, thereby magnifying the case's influence on parameter estimates, both the fixed-effects and the random components. Therefore, in this subsection, we discuss some basic diagnostic techniques to identify influential observations in LMMs.

Cook's D and DFBETAS Statistic

In fitting an LMMs, some observations may have unduly impacted the inferential process to derive parameter estimates. Conventionally, these influential cases can be identified by the change in the estimated regression coefficients after deleting each observation in a sequence (Liu, 2015; Cook, 1977, 1979). As noted by Liu (2015), for a single covariate x_m, the distance in the estimated regression coefficient after removing the subject denoted $\hat{d}_{m_i}$, is written as

$$\hat{d}_{m_i} = \hat{\beta}_m - \hat{\beta}_{m_{(-i)}},$$

where $\hat{\beta}$ represents the estimate of β from the full data and $\hat{\beta}_{m_{(-i)}}$ be the estimate of β after subject i has been eliminated from the data.

This statistic can be expressed in terms of the entire vector of the estimated regression coefficients, given by $\widehat{d}_i = \hat{\beta} - \hat{\beta}_{(-i)}$. A greater value of $\widehat{d}_i$ suggests subject i to have a stronger influence on the estimate of β; likewise, a lower value indicates that subject i impact on the model fit is limited (Liu, 2015).

For LMMs, the specification of the scaled Cook's D is given by

$$\bar{d}_i = \frac{[\hat{\beta} - \hat{\beta}_{(-i)}]' Var(\hat{\beta})^{-1} [\hat{\beta} - \hat{\beta}_{(-i)}]}{rank(x)},$$

where $Var(\hat{\beta})$ represents the covariance matrix after a case has been estimated from the data (Christensen et al., 1992). As an alternative to cook's D, DFBETAS statistic is useful to identify the change in the parameter estimates by influential observations, defined as

$$DFBETAS_i = \frac{\hat{\beta}_i - \hat{\beta}_{-ij}}{s.e(\hat{\beta}_{-ij})},$$

where i represents the parameter estimates in the j^{th} group, $\hat{\beta}_i$ and $\hat{\beta}_{-ij}$ are respectively, the estimate base of the full data and the estimate after eliminating group j from the data. Furthermore, $s.e(\hat{\beta}_{-ij})$ is the standard error of $\hat{\beta}_{-ij}$.

Cook's D measures the influence of a single level group on all parameter estimates (Nieuwenhuis et al., 2012). However, according to Fox & Monette (2002), DFBETAS quantifies the influence of observations on single parameter estimates. With this in mind, the DFBETAS is essential in evaluating the reliability of specific estimates individually. On the other hand, Cook's D is highly valuable for assessing the reliability of all group-level estimates simultaneously (Van der Meer et al., 2010). As a criterion, Belsley et al. (2005) conclude that cases are regarded as influential if the associated absolute value of DFBETAS or Cook's distance values exceeds the cut off value, $2/\sqrt{n}$ or $4/n$ respectively, with n representing the number of groups in the grouping variable.

Likelihood Distances

An overall influence statistic measures the change in the objective function being minimized (Schabenberger, 2005). In LMMs, fit by ML or REML, which are the two likelihood-based methods implemented in the SAS Proc Mixed package, the overall influence measure is the likelihood distance (Schabenberger, 2005). The reduced log-likelihood function ℓ and restricted log-likelihood function ℓ_R of the LMM is given as follows:

$$ML : \ell = -\frac{1}{2} log|\boldsymbol{V}| - \frac{1}{2} \boldsymbol{r}' \boldsymbol{V}^{-1} \boldsymbol{r} - \frac{n}{2} log(2\pi),$$
$$REML : \ell_R = -\frac{1}{2} log|\boldsymbol{V}| - \frac{1}{2} log|\boldsymbol{X}' \boldsymbol{V}^{-1} \boldsymbol{X}| - \frac{1}{2} \boldsymbol{r}' \boldsymbol{V}^{-1} \boldsymbol{r} - \frac{n-p}{2} log(2\pi),$$

where $\boldsymbol{r} = \boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X'V^{-1}X})^{-}\boldsymbol{X'V^{-1}Y}$, and p is the rank of X. The likelihood and restricted likelihood distances, denoted by LD_u and RLD_u, respectively, can then be defined as

$$LD_u = 2\{\ell(\hat{\psi}) - \ell(\hat{\psi}_u)\},$$
$$RLD_u = 2\{\ell_R(\hat{\psi}) - \ell_R(\hat{\psi}_u)\},$$

where $\ell(\hat{\psi})$ represents the full data parameter estimates (collection of all the fixed effects $\boldsymbol{\beta}$ and the covariance parameters $\boldsymbol{\theta}$), and $\ell(\hat{\psi}_u)$ represents the reduced data parameter estimates.

Covariance Ratio (CovRatio)

A determinant operation such as CovRatio is one of the common ways to do influence on the precision of estimates. The covariance-based statistics measure the impact on the precision of estimates, whereas Cook's D measures the impact of data points on the parameter estimates (Littell et al., 2006). The SAS Mixed Procedure computes CovRatio of the fixed-effect parameters as follows:

$$CovRatio(\boldsymbol{\beta}) = \frac{det_{ns}(\hat{Var}[\hat{\boldsymbol{\beta}}_u])}{det_{ns}(\hat{Var}[\hat{\boldsymbol{\beta}}])},$$

For covariance parameter estimates:

$$CovRatio(\boldsymbol{\theta}) = \frac{det_{ns}(\hat{Var}[\hat{\boldsymbol{\theta}}_u])}{det_{ns}(\hat{Var}[\hat{\boldsymbol{\theta}}])},$$

where $det_{ns}(M)$ represents the determinant of the nonsingular part of matrix M (Christensen, 1991; Littell et al., 2006).

The covariance ratio statistic relates the determinants of the variance matrices of the full-data and reduced-data estimates. The measure of no influence is a value of 1. Values larger than 1 indicate increased precision in the full-data case, and values smaller than 1 indicate higher precision for the reduced-data estimates (Littell et al., 2006).

Predicted Residual Sum of Squares (PRESS)

In addition to analyzing a unit's influence on the change in parameter estimates, and the change within the precision of estimates, influence on fitted and predicted values can also be inspected through the PRESS statistic. The PRESS provides a comparison of the predicted marginal mean and the observed mean when the predicted value is

calculated without the deleted observation in question (Schabenberger, 2005). The PRESS statistic is the sum of the squared PRESS residuals,

$$PRESS = \sum_{i \in u} \hat{\varepsilon}_{i(u)},$$

where the sum is over the observations in u.

In general, "influence" is understood as the capability of a single or multiple data points, via their presence or absence in the data, to alter important aspects of the analysis, yield qualitatively different inferences, or violate assumptions of the statistical model. The primary goal of influence analysis is not to mark data points for deletion so that a good model fit can be achieved for the reduced data. However, this might be a result of the influence analysis. The goal is instead to determine which cases are influential and how they are essential to the study (Schabenberger, 2005). It is vital to note that the influence analysis is performed under the assumption that the chosen model is correct. Changing the model structure can alter the conclusion.

2.6 Data example: CAPRISA 002 Acute Infection Study

This section illustrates the estimation, methodology, and model selection procedures discussed in the above sections on the Centre for the AIDS Programme of Research in South Africa (CAPRISA) 002 Acute Infection Study data set. Between August 2004 and May 2005, CAPRISA initiated a cohort study enrolling high-risk HIV-negative women to follow up. These women then followed up closely to study disease progression and CD4/viral load evolution (Garrett et al., 2018; Mlisana et al., 2014; Moosa et al., 2018). The study was conducted at the Doris Duke Medical Research Institute (DDMRI) at the Nelson R Mandela School of Medicine of the University of KwaZulu-Natal in Durban, South Africa. This study observed $N = 235$ incident HIV-1 positive women whose disease biomarkers were (CD4 counts and Viral Loads) measured repeatedly at least four times on each participant.

The baseline characteristics of the dataset are given in Table 2.3. From a total of 235 women, 105 (44.7%) were residing around Vulindlela (rural site), and 130 (55.3%) were living around eThekwini (Durban, urban site), KwaZulu-Natal, South Africa. The average age at enrollment and baseline square root transformed CD4 cell counts were 27.15 years (range 18-59) with a standard deviation of 6.56 and 23.45 (range 13-40) with a standard deviation of 4.594, respectively. The average follow-up time was 2.69 years, and the majority of the women, 182 (77.4%), had a stable partnership. Furthermore, from the total women included in the study, the majority of the 224

(95.3%) completed secondary/high education, and most of the women (78.8%) were self-reported sex workers (Mlisana et al., 2014; Van Loggerenberg et al., 2008). There were a total of 7129 observations from the 235 women, which consists of a minimum of two and a maximum of sixty-one measurements of CD4 cell counts, among the subjects which were measured at different time points, indicating that the number of measurements over all subjects was not equal. Further apart from an unequal number of measurements across individuals, measurements were not taken at fixed time points, which implies the CAPRISA 002: Acute Infection Study is a highly unbalanced longitudinal data set that requires carefully designed modeling approaches.

Table 2.3: Baseline characteristics of the CAPRISA 002 AI Study data set, 2004-2018

Variable	Total	Variable	Total
Number of women	235	**Marital Status**	
Place of residence		No partner	43 (18.3%)
Rural	105 (44.7%)	Stable partner	182 (77.4%)
Urban	130 (55.3%)	Many partners	10 (4.3%)
Age at Seroconversion (Years)			
Mean (Std. Deviation)	27.15 (6.56)	**Educational Level**	
$\leq$20	21 (8.9%)	Primary schools (grade 0-7)	11 (4.7%)
20-29	150 (63.8%)	Secondary schools (grade 8-12)	224 (95.3%)
30-39	50 (21.3%)	**Baseline sqrt of CD4 cell counts (cells/ML)**	
40-49	12 (5.1%)	Mean (Std. Deviation)	23.5 (4.594)
$\geq$ 50	2 (0.9%)	**Baseline HIV viral load (cells/μL)**	
Baseline Body Mass Index		Undetectable VL ($\geq$ 50)	1 (0.4%)
Underweight	14 (6%)	Low VL ($50 \leq VL \leq 10000$)	74 (31.5%)
Normal weight	173 (73.6%)	Medium VL ($10000 \leq VL \leq 100000$)	94 (40%)
Overweight	41 (17.4%)	High VL ($\geq$ 100000)	66 (28.1%)
Obese	7 (3%)		

Figure 2.1 (left panel) shows that CD4 cell count distribution is right-skewed, indicating non-normality; thus, a square root transformation to CD4 cell count was performed to normalize the data, Figure 2.1 (right panel) shows that the square root transformed data conforms quite well to the normal distribution.

The spaghetti plots in Figure 2.2 illustrate the actual CD4 cell count measurements for randomly chosen participants over time across pre and post ART initiation groups. Since plots with all individual curves can be hard to distinguish for a large sample

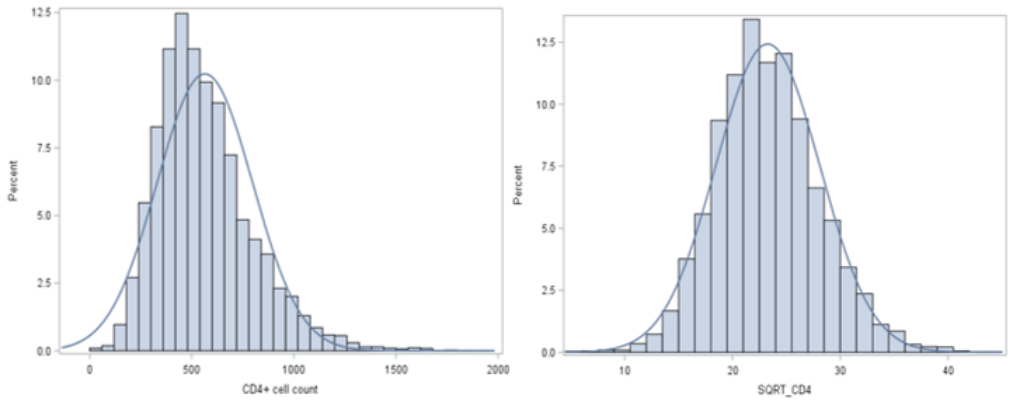


Figure 2.1: Distributional properties plot for original and square root transformed CD4 trajectories

size, we randomly chose 15 participants to construct such individual plots. Figure 2.2 shows a decreasing trend of CD4 cell count over time on patients before Highly Active Antiretroviral Therapy (HAART) initiation, but an increasing trend of CD4 cell count over time for the same 15 randomly chosen patients initiated on HAART. Figure 2.2 also indicates that there is evidence of variability between individuals as well as variability within individuals. In addition, the individual profiles are not all of the same lengths, an indication of incompleteness and missing data due to dropout or attrition.

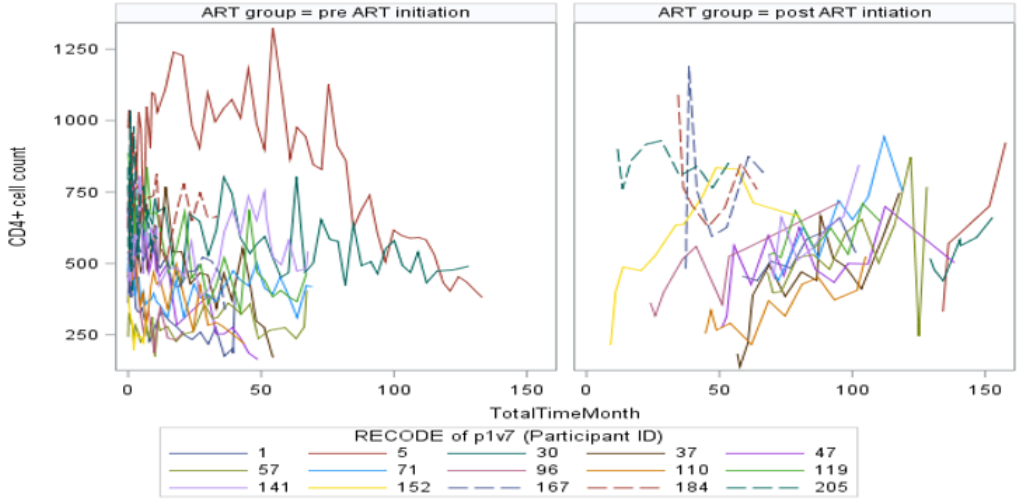


Figure 2.2: Individual profiles plot of CD4 count for the same 15 randomly selected individuals before and after HAART

Figure 2.3 shows an array of individual series from the CAPRISA 002: AI study. In each panel, the observed CD4 count for a single subject is plotted against the times

that measurements were obtained. Such plots permit assessment of the person response patterns and whether there is substantial heterogeneity within the trajectories. Figure 2.3 shows that there can be variation in the "level" of CD4 count for subjects. Subject PID=5 in the first row second from left has CD4 counts greater than 500 for almost all times while PID= 110 in the third row lower-left corner has all measurements below 500. Moreover, PID=30 in the first row third from left has all measurements almost constant around 500. Further, individuals profile plots can be evaluated for the change over time (Twisk, 2013). Figure 2.3 shows that most subjects are either relatively stable in their measurements over time or tend to be increasing.

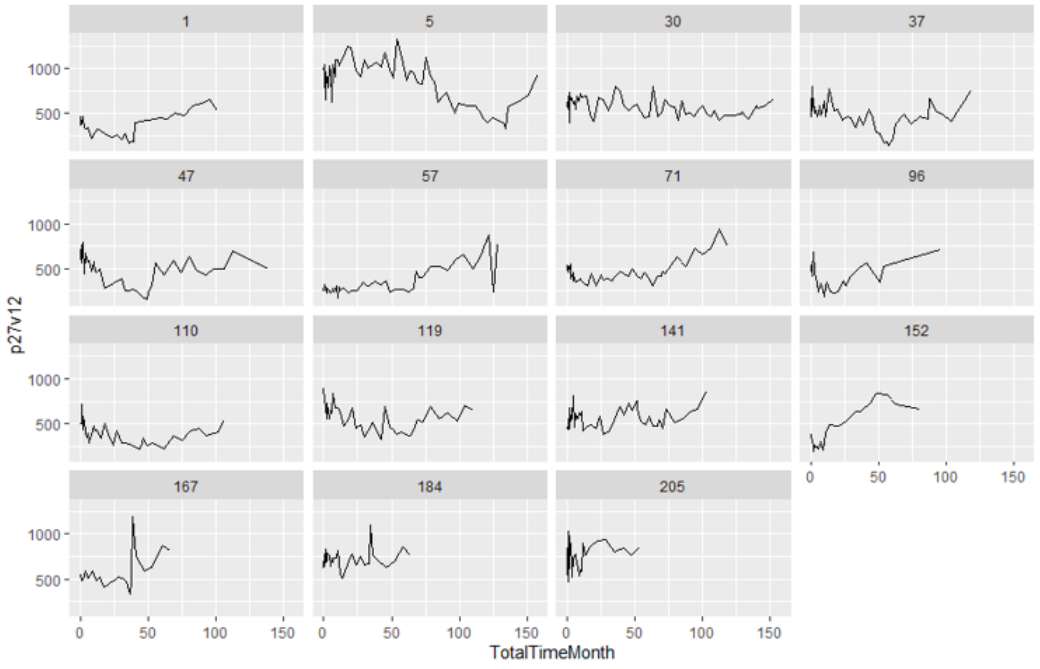


Figure 2.3: A sample of 15 individual CD4 trajectories from the CAPRISA 002 AI Study

Figure 2.4 shows the mean CD4 trajectories over time for the pre and post ART initiation group in the CARISA 002: AI study. Overall the mean plots suggest that patients initiated on HAART have significant quadratic growth in the evolution of CD4 count over time as what we would expect. Furthermore, the plots appeared to be nonlinear implying factor that controls the nonlinear effect that may need to be applied to the data.

The inferential focus of this study is on the mean response of a square root transformation to the CD4 cell count measure. An appropriate selection of the random effects was also performed. That is, the appraisal as to which of the nonlinear components (the intercept, time, or square root of time) ought to have a random component was made. A covariance structure must be incorporated into the statistical

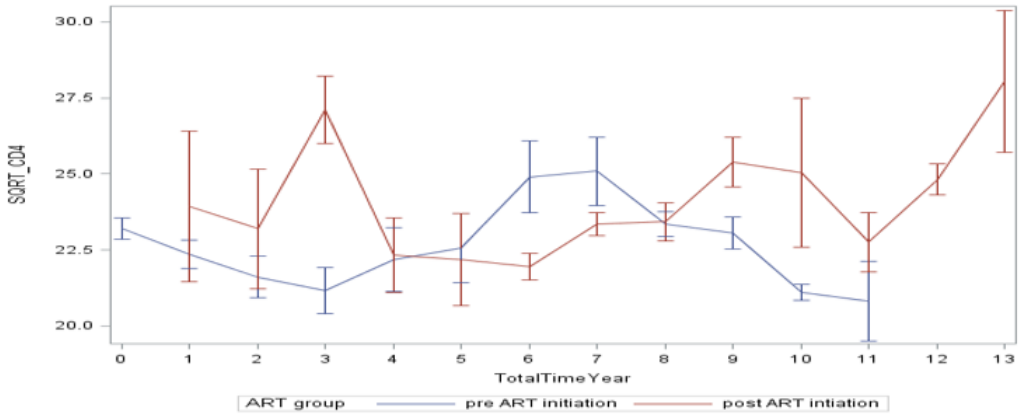


Figure 2.4: Mean CD4 trajectories over time by ART Initiation group, CAPRISA 002 AI study

model to have a valid inference about the mean structure (Melesse & Zewotir, 2017). Hence, following the selection of random components, a comparison of covariance structure was made in the study.

The following random effect models, which have the same fixed effects, were fitted for testing:
Model 1: Intercept, Time, SQRT of time (*Random intercept and slope (time and SQRT of time) model*)

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 t_{ij} + \beta_2 \sqrt{t_{ij}} + b_{i0} + b_{i1} t_{ij} + b_{i2} \sqrt{t_{ij}} + \varepsilon_{ij}$$

where x_{ij} is the ART initiation group variable, and t_{ij} is the time variable.

Model 2: Time, Square root of time (*Random slope model*)

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 t_{ij} + \beta_2 \sqrt{t_{ij}} + b_{i1} t_{ij} + b_{i2} \sqrt{t_{ij}} + \varepsilon_{ij}$$

Model 3: Time only (*Random slope model without random SQRT of time*)

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 t_{ij} + \beta_2 \sqrt{t_{ij}} + b_{i0} + b_{i1} t_{ij} + \varepsilon_{ij}$$

Model 4: Intercept only (*Random intercept model*)

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 t_{ij} + \beta_2 \sqrt{t_{ij}} + b_{i0} + \varepsilon_{ij}$$

All models were fitted using the REML estimation procedure, and model comparison is made using different Information Criteria. Hence, we conclude that the random intercept and slope model is the preferable model among the models listed

Table 2.4: Model comparison using IC for random effects using REML estimation

		Information Criteria				
Random effect models	Params	-2log ℓ	AIC	HQIC	BIC	CAIC
Model 1	**4**	**34392.7**	**34400.7**	**34406.3**	**34414.6**	**34418.6**
Model 2	3	36567.8	36573.8	36577.9	36584.1	36587.1
Model 3	2	39832.4	39836.4	39839.2	39843.3	39845.3
Model 4	2	36363.7	36367.7	36370.5	36374.6	36376.6

above (Table 2.4).

To validate the random intercept and slope model, a panel of conditional studentized residuals for the square root CD4 count was used. The result is presented in Figure 2.5. The panel consists of a scatterplot of the residuals, a histogram with normal density, a Q-Q plot, and summary statistics for the residuals and the model fit. The residuals were randomly dispersed around zero, suggesting that their mean was approximately zero. The histogram follows a normal distribution indicating a constant variance, which was moreover affirmed by the Q-Q plot that did not show heavy tails. Hence, the fulfillment of the assumption that the error term ϵ_{ij} was normally distributed with mean 0 and variance σ^2.

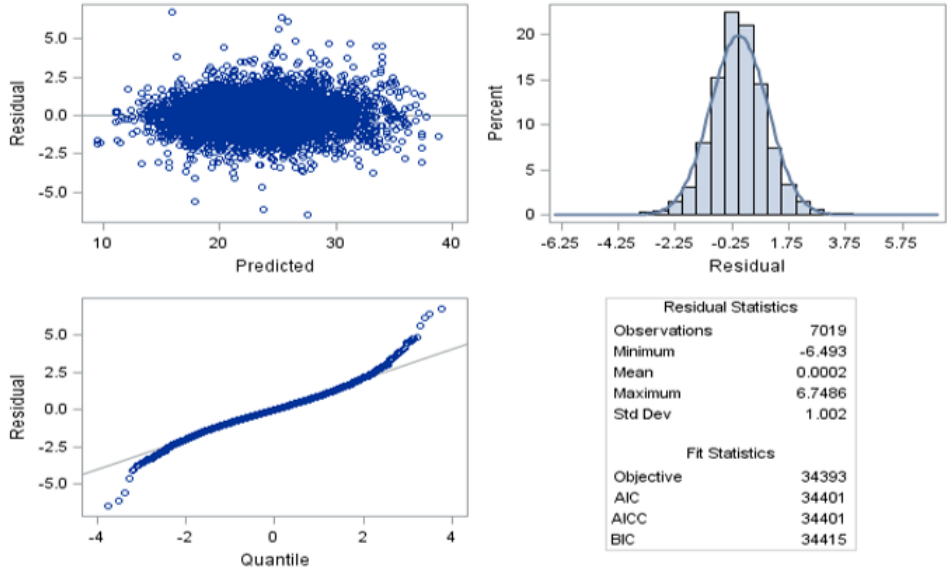


Figure 2.5: Panel of conditional studentized residuals for the square root of CD4 count

Table 2.5 shows the comparisons between the four different covariance structures

that were considered in the model using REML under the same fixed effects model. The Information Criteria were used to compare models for the structure that gives a better fit.

Table 2.5: Comparisons of covariance structure

		Information Criteria				
Covariance Structure	Params	-2log ℓ	AIC	HQIC	BIC	CAIC
AR(1)	3	35675.6	35681.6	35685.8	35692.0	35695.0
CS	3	35671.5	35677.5	35681.7	35687.9	35690.9
TOEP	4	35671.4	35679.4	35685.0	35693.2	35697.2
UN	7	**34087.1**	**34101.1**	**34110.8**	**34125.3**	**34132.3**

The estimated unstructured covariance parameter determines the matrix $(\hat{G})$ along with the estimated variance of the random error term $(\hat{R})$, respectively, are given below for Model 1:

$$\hat{G} = \begin{bmatrix} 20.1224 & 0.09786 & -2.4719 \\ 0.09786 & 0.01849 & -0.1705 \\ -2.4719 & -0.1705 & 1.9686 \end{bmatrix} \qquad and \qquad \hat{R} = var(\epsilon_{ij}) = 5.7063$$

Table 2.6 shows the REML estimates for the fixed effects of the random intercept and slope model (Model 1).

Table 2.6: Fixed effect estimates of Model 1 for unstructured covariance structure

Effect	DF	Estimate	SE	Pr $< \lvert t \rvert$	95% C.I for Estimate
Intercept	234	24.3062	0.3055	$<$.0001	(23.7043, 24.9081)
Time in month	6781	0.09015	0.01072	$<$.0001	(0.06913, 0.1112)
Sqrt_Time	6781	-0.9554	0.1036	$<$.0001	(-1.1586, -0.7523)
ART Initiation (Post)	195	2.4473	0.1348	$<$.0001	(2.1815, 2.7131)

The overall mean CD4 count for post ART initiation group is 26.7535, whereas the mean at time $'t'$ is estimated as

$$\hat{Y}_t = 26.7535 + 0.09015t - 0.9554\sqrt{t}$$

The overall mean CD4 count for pre ART initiation group is 24.3062, whereas the mean at time 't' is estimated as

$$\hat{Y}_t = 24.3062 + 0.09015t - 0.9554\sqrt{t}$$

The above fitted conditional models are extended to incorporate the impact of patient's age, educational status, number of sex partners, baseline BMI, baseline viral load, and ART initiation group with the square root of CD4 count as the response. In addition to this, the two-way interaction effect was evaluated within the modeling process. But, none of the interaction effects was significant. The results of the effects of age, educational status, and the number of sex partners were not found to be significant. However, we incorporate them within the modeling process since factors with subject matter importance ought to be kept within the model to eliminate any confounding effects.

Table 2.7: Fixed effect estimates of the full Model

Covariates	Estimate	SE	Pr $< \|t\|$	95% C.I for Estimate
Intercept	25.2439	0.6040	$<.0001$	(24.0536, 26.4342)
Time in month	0.06377	0.009142	$<.0001$	(0.04585, 0.08169)
Sqrt_Time	-0.6674	0.09020	$<.0001$	(-0.8442, -0.4906)
ART Initiation (Post)	2.1104	0.1647	$<.0001$	(1.7855, 2.4353)
Baseline BMI category (ref.=Normal weight)				
Obese	8.0201	1.2896	$<.0001$	(5.4788, 10.5614)
Overweight	0.4966	0.5799	0.3927	(-0.6461, 1.6394)
Underweight	0.2486	0.9131	0.7856	(-1.5508, 2.0481)
Baseline HIV viral load category (ref.= Low VL)				
High VL	-3.2552	0.5633	$<.0001$	(-4.3652, -2.1452)
Medium VL	-1.5696	0.5211	0.0029	(-2.5965, -0.5426)
Undetectable	1.3418	3.3359	0.6879	(-5.2321, 7.9157)
Number of sex partner (ref.= Stable partner)				
Many partners	-1.4706	1.0859	0.1770	(-3.6105, 0.6693)
No partner	-0.6478	0.5791	0.2645	(-1.7889, 0.4933)
Age group (ref.= < 20)				
20-29	0.06144	0.4231	0.8847	(-0.7742, 0.8971)
30-39	0.1611	0.4780	0.7366	(-0.7831, 1.1053)
40-49	0.2491	0.6420	0.6985	(-1.0190, 1.5172)
50-59	-1.0100	1.0149	0.3212	(-3.0147, 0.9946)
≥ 60	-0.7631	1.9554	0.6969	(-4.6254, 3.0991)
Education attainment (ref.= Secondary or high school)				
Primary school	0.08077	1.0585	0.9392	(-2.0052, 2.1668)
Residence of participant (ref.= Urban)				
Rural	-0.2647	0.4539	0.5604	(-1.1593, 0.6298)

The results of the fixed effect estimates are presented in Table 2.7. As seen from Table 2.7, the model intercept ($\hat{\beta}_0$) is equal to 25.2439, which is an estimate of the mean square root CD4 count at baseline (i.e., month=0) subject to other effects with covariate values set to zero in the model. The Month effect ($\hat{\beta}_1$)=0.06377 is the slope or rate of change in the mean square root CD4 count per unit increase in the month among HIV-infected patients with other covariate values set to zero. In other words, the time (month) effect shows a significant positive effect on the mean CD4 count with a rate of 0.06377 (p-value <0.0001) units per month. Hence square root of CD4 count increases by 0.06377 for every month among patients, showing low progress of CD4 count over time. The effect of the square root of time (p-value $<$ 0.0001) is also significant but appears to have an opposite effect on the square root of CD4 count in a cohort of HIV-infected patients enrolled in the CAPRISA 002 Acute Infection Study. The estimate for post-HAART initiation shows a highly significant positive effect with a mean square root CD4 count of 2.1104 units higher than the pre-HAART state. This implies, among patients in the post-HAART initiation group, their mean square root CD4 count increased by 2.1104, but this is not a slope. Relative to patients with normal weight status, patients with higher BMI (Obese) show a highly significant positive effect (p-value<0.0001) with 8.0201 square root CD4 count higher than the reference group (Table 2.7). However, underweight patients (patients with low BMI) show no significant effect compared to the reference group. After the patients had been initiated on HAART, the average square root CD4 count among patients with a high value of the viral load at baseline is -3.2552 (p-value<0.0001) units lower compared to patients with low viral load at baseline. Moreover, after the patient had been initiated on HAART, the average square root CD4 count among patients with a medium viral load category at baseline is decreased by 1.5696 (p-value=0.0029) units compared to the average square root of CD4 count among patients with low viral load at baseline. Implying that patients with high and medium viral load at baseline have significantly lower mean CD4 count compared to patients with low viral load at baseline.

Table 2.8 shows a comparison of the three commonly used spatial covariance structures: spatial exponential structure (SP(EXP)), spatial spherical structure (SP(SPH)), and spatial Gaussian structure SP(GAU). Since the exponential model has the smallest information criteria statistics and the smallest $-2log\ell$ suggests that the SP(EXP) structure is the best of the three spatial covariance models (Table 2.8).

The estimated spatial exponential covariance parameters are demonstrated in Table 2.9. The estimate of the *sill* (σ^2) is 9.7063, reported as "Variance", which corresponds to the variance of observation (Table 2.9). The estimated *range* ($\rho(h)$) is 31.1376,

Table 2.8: Comparison of spatial covariance models

	Model Fitting Criteria				
Spatial covariance	Params	-2log ℓ	AIC	HQIC	BIC	CAIC
SP(EXP)	**9**	**33024.5**	**33042.5**	**33055.1**	**33073.6**	**33082.6**
SP(SPH)	9	33039.1	33057.1	33069.6	33088.2	33097.2
SP(GAU)	9	33162.1	33180.1	33192.7	33211.2	33220.2

which appears as "SP(EXP)," which is the practical range or distance at which the spatial autocorrelation in the exponential model is three times this amount, $3 \times 31.1376 = 93.4128$. That is, observations separated by more than 93.4128 distance units are not spatially correlated. The estimated *nugget* (C_0) is 3.4986, which appears as "Residual," that is, the value at which h=0 or defined as "Intercept" in the spatial covariance structure model.

Table 2.9: Covariance Parameter Estimates of the full model

Cov Parm	Estimate	SE	Z Value	Pr>Z
UN(1,1)	3.3317	2.6772	1.24	0.1067
UN(2,1)	0.05870	0.04370	1.34	0.1792
UN(2,2)	0.004944	0.001733	2.85	0.0022
UN(3,1)	-0.3405	0.4031	-0.84	0.3983
UN(3,2)	-0.05410	0.01654	-3.27	0.0011
UN(3,3)	0.6223	0.1798	3.46	0.0003
Variance	9.7063	2.3528	4.13	<.0001
SP(EXP)	31.1376	9.4724	3.29	0.0005
Residual	3.4986	0.1008	34.70	<.0001

Figure 2.6 indicates the predicted profile plot for the average number of CD4 cells, following Table 2.6 shows results obtained by the fitted mixed-effects model. The predicted values closely matched the observed CD4 count, with an $R^2 = 0.75$, suggested that the overall model fit was good (Figure 2.7).

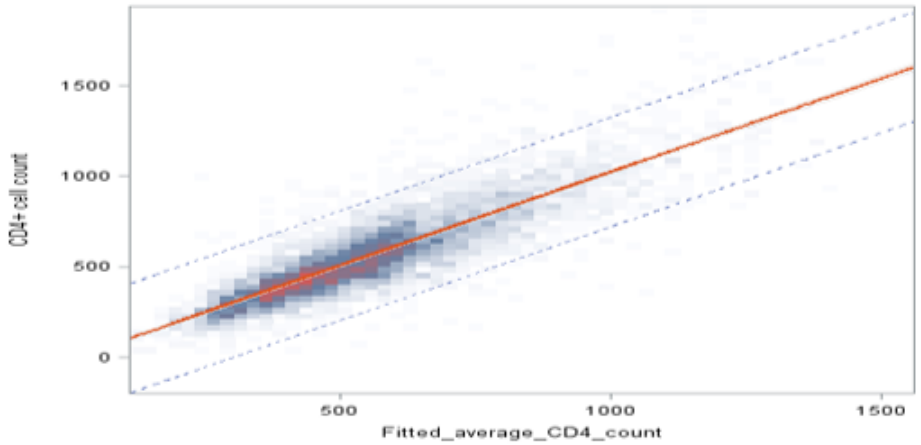


Figure 2.6: Heat map of fitted average by observed CD4 count overlaid with the fitted line

The fitted solid line in the above figure also indicates the estimated regression line between the observed CD4 count and fitted values (Fitted= 148.07+0.7259 observed), and the two dashed lines show both 95% confidence interval and prediction interval.

The overall influence diagnostic and diagnostics for the fixed effects are displayed graphically hereunder in Figure 2.7-2.11. Figure 2.7 shows the needle plot of the Restricted Likelihood Distance (RLD) for the response variable (square root of CD4 count). The RLD plot suggests that the overall influence of patients 5, 12, 29, 32, 55, 84 and 188 stands out compared to those of the rest of the patients (Figure 2.7).

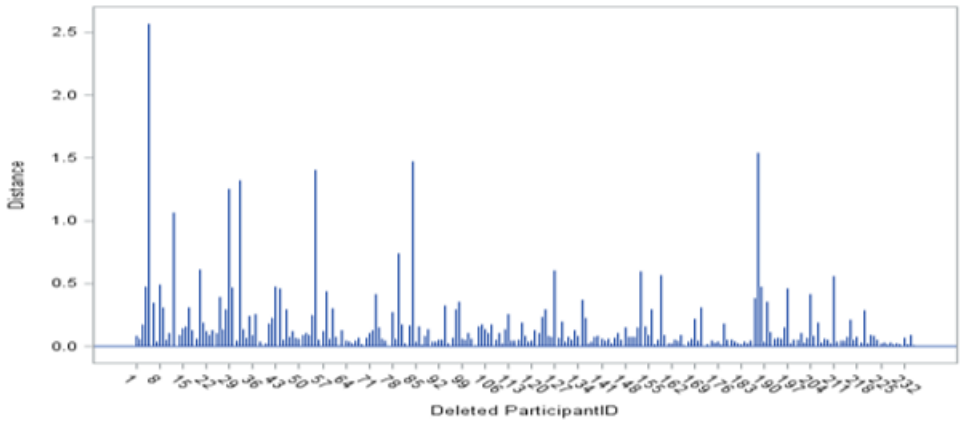


Figure 2.7: Restricted Likelihood Distance

PRESS statistics are sums of squared PRESS residuals in the deletion sets (Schabenberger, 2005). Figure 2.8 shows the scatter plot of the PRESS statistics for the square root of the CD4 count. Large values of the PRESS statistic for patients 5, 60, 84, 127, and 189 are noted.

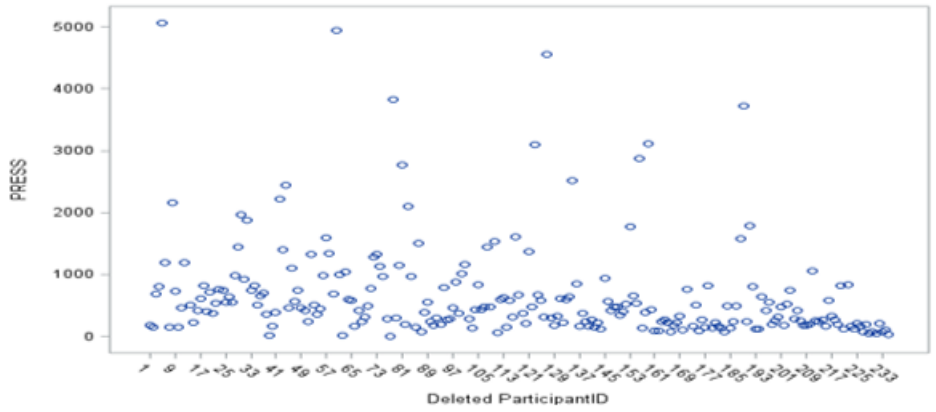


Figure 2.8: PRESS Statistics

A panel of influence statistics for fixed effects and covariance parameters is presented in Figure 2.9. Cook's D statistics measure the influence on the vector of parameter estimates, and the CovRatio statistic measures influence on the covariance matrix of the parameter estimates. The patients with the most substantial effect on the fixed effect estimates are 5, 32, and 55 (Cook's D Fixed effects). Cook's D Covariance parameters indicate that the influence of patients 12, 84, and 188 far exceeds that of other subjects in the study data sets. This is expected since their RLD is substantial, while their impact on the fixed effects was relatively moderate. The CovRatio Covariance Parameters also show that the covariance parameters may be estimated much more precisely in the absence of those patient's observations, especially patients 84 and 188. Note that other sets of observations, besides those listed above, exert influence on the chosen model (Model 1).

A panel of deletion estimates for the response variable is displayed in Figures 2.10 and 2.11 to examine how the individual parameter estimates and covariance parameters, respectively, react to the removal of the influential sets of observations (Schabenberger, 2005). Each cell in the panel (Figure 2.10) displays the estimates of few fixed effects that were included in the fitted model. Each cell in Figure 2.11 displays estimates of the 3×3 variance-covariance matrix of the random coefficients and the estimate of SP(EXP) parameter following removal of sets of influential observations. Reference lines are drawn at the complete-data parameter estimates.

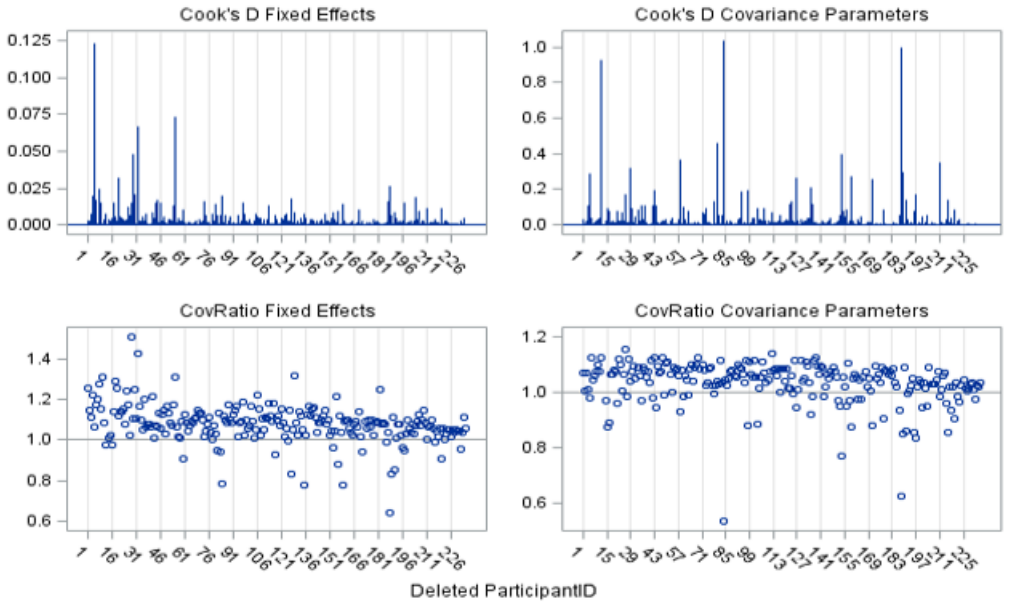


Figure 2.9: Influence statistics for the square root of CD4 count

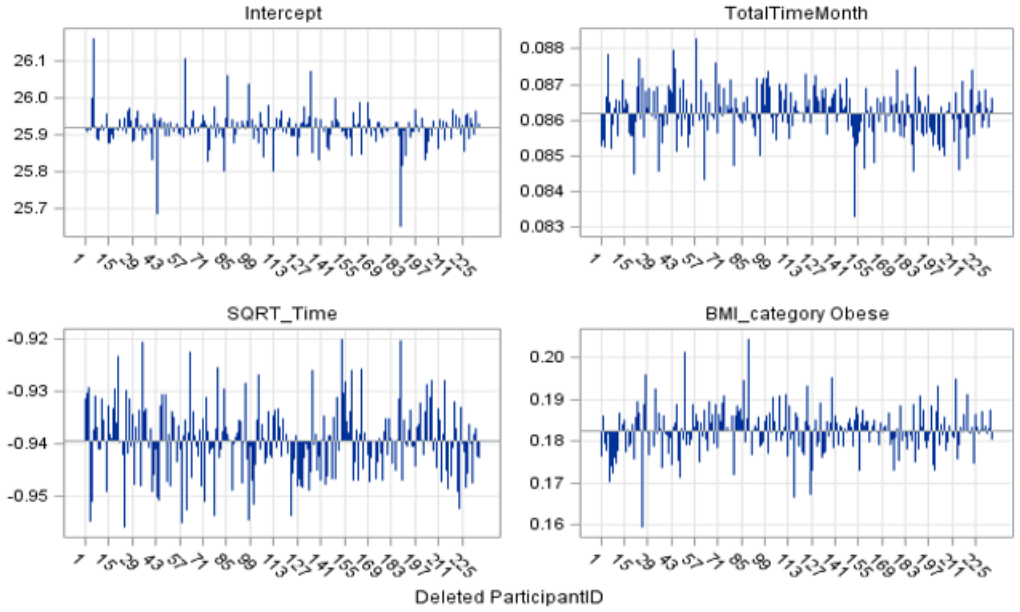


Figure 2.10: Fixed effects deletion estimates for square root of CD4 count

The focus of Figure 2.10 is on the behavior of individual parameter estimates that react to the removal of influential cases. Specifically, subjects 5, 44, 60, and 188 indicate a substantial impact on the model fit of the intercept. However, the removal of these subjects does not influence at all the displayed fixed effects. On the other hand, subject 27 is identified as an additional influential case since it strongly impacts the obese BMI category (Figure 2.10). Subjects 5, 29, 73, and 85 are also identified as influential cases since their presence in the data reduces the estimate of SP(EXP) parameter (Figure 2.11), substantially reducing the degree of correlation among data points from any patient. On the other hand, observation from subject 12 has the opposite effect. The temporal correlation drops when the impact of this patient's data is removed.

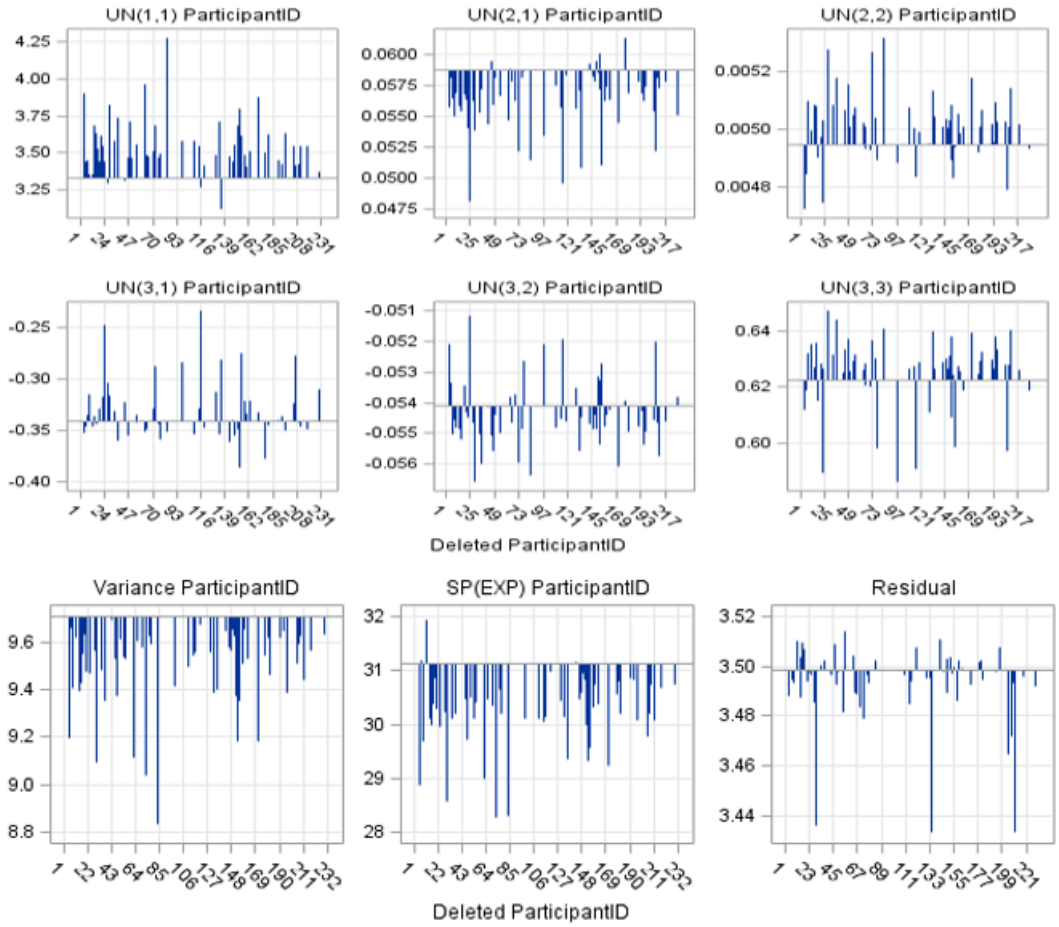


Figure 2.11: Covariance parameter deletion estimates for square root of CD4 count

Finally, the normal probability plot of the random effects for the fitted mixed-effects model is indicated in Figure 2.12. The assumption of normality seems reasonable for all three random effects (Pinheiro & Bates, 2006).

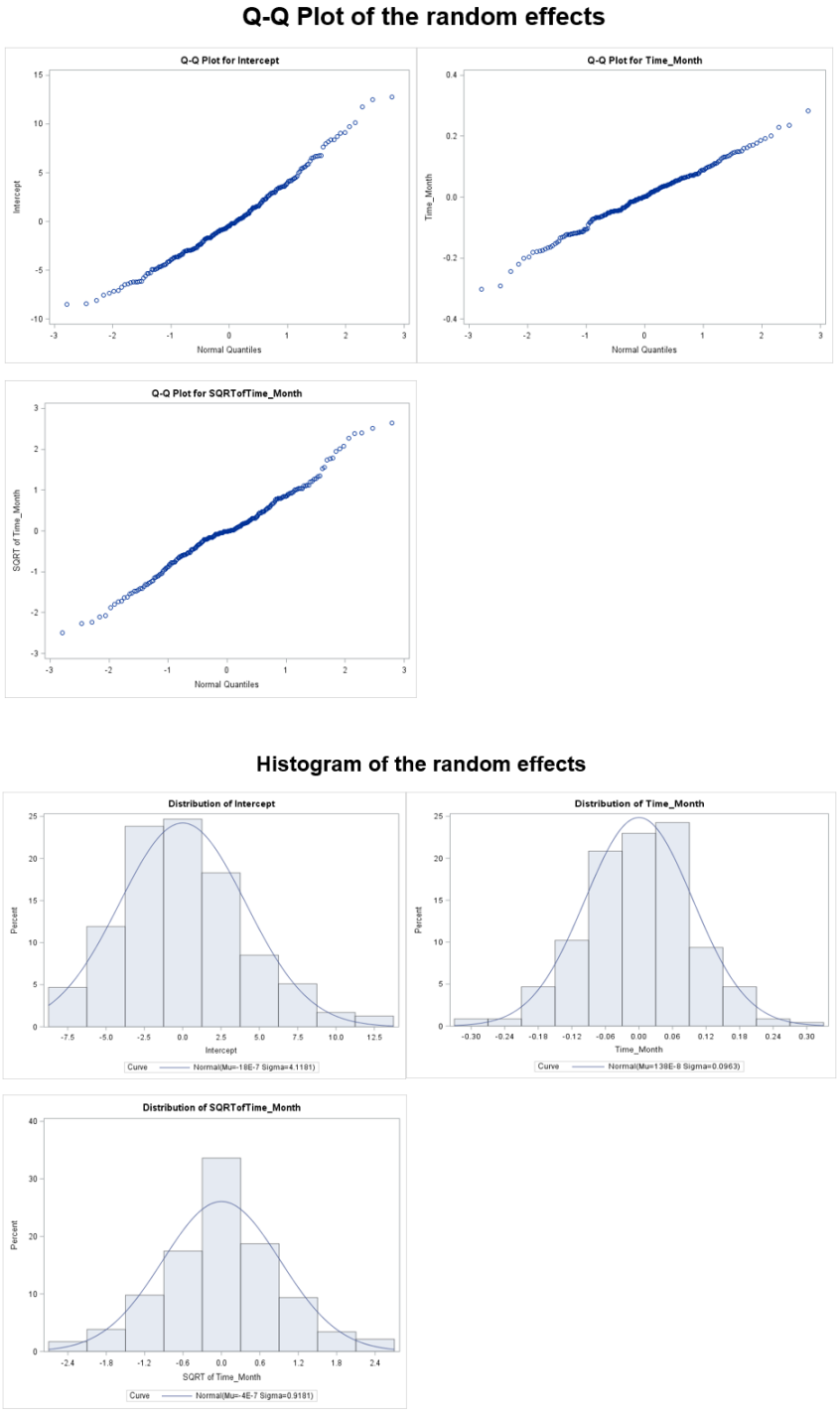


Figure 2.12: Q-Q and Histogram normal plot of estimated random effects

Some of the codes that are used for this section can be found here (Code 7.1 in the Appendix A).

2.7 Summary

Mixed models are one of the special statistical models that are useful in understanding longitudinal or repeated measures data. The models permit the examination of the changes over time within and between subjects. In the presence of fixed effects and random effects, the selection of an appropriate mixed model is more complicated than for a linear regression model. The fixed effect and the random effect structure are subordinate to each other and the determination of one influences the other (Melesse & Zewotir, 2017). In this study, a step-up model selection procedure was applied to find a reasonable model that fits the data, primarily since this procedure begins with the simplest possible model and is built up by including more covariates within the model and hence does not have much numerical issue (West et al., 2014; Melesse & Zewotir, 2017; Diggle et al., 2002). In this study, the model where the intercepts and slopes were considered as random effects consolidated with the UN covariance structure. The fixed effects combined with the REML estimation technique were determined as the best fit to estimate the prognosis of the square root transformation of the CD4 count of HIV-infected patients enrolled in the CAPRISA 002 Acute Infection Study.

The results revealed that the prognosis of the CD4 count of a patient is significantly increased after the patient had been initiated on HAART, as we would anticipate. The impact of HIV-infected patients with the predominance of obese nutrition status (higher BMI) at baseline showed significance after patients had been initiated on HAART. Therefore, we ought to pay more consideration to the BMI of HIV-infected patients before and after HAART initiation. This may inform future techniques in studying the progression and the immunologic responses to treatment, but that does not infer that patients with higher BMI ought to be clinically ignored. Instead, based on this study and other findings, it appears that BMI contributes to some degree to drug metabolism and consequently influencing the proficiency of HAART (Palermo et al., 2011; Li et al., 2019). Moreover, our results also showed that the impact of patients with higher viral load before the patient had been initiated on HAART significantly reduced their CD4 count. Therefore, effective HAART initiation after HIV exposure is necessary to suppress the increase of viral loads to induce potential ART benefits that accrue over time.

The results of the influence diagnostics analysis for the CAPRISA 002 Acute Infection study using the chosen mixed-effects model were also performed. Several cases were identified as influencing the analysis of the fitted model. Influence diagnostics analysis is essential for statistical analysis to determine how individual observations

or sets of observations are influential that their presence or absence from the data impacts the analysis (Zewotir, 2008). The goal of influence analysis is not to determine observations for removal from the analysis but to determine which cases exert undue influence on the analysis. Eliminating certain subjects from the data and basing the final analysis on only the remainder is usually not the right action to take. The results of a diagnostic influence analysis can be seen only in light of the model we are working with (Littell et al., 2006).

Moreover, the data showed evidence of strong individual-specific effects on the evolution of CD4 counts. The diagnostic plots also suggested a significant individual heterogeneity between subjects both before and after HAART initiation. Thus, this may indicate that prescribing a common treatment or intervention overall to patients may not be the best strategy. More research may be required to understand what factors cause patients to respond differently to treatment intervention. Such information may help design treatment and intervention strategies that may be more efficient to a specific group of patients rather than one treatment/intervention fit all strategy.

The models depicted in this study may empower the description of the effect of several covariates on the square root CD4 count of HIV-infected patients utilizing all accessible information. We believe that this sort of analysis can be valuable to address several important issues in public health as well as offer assistance in observing patients and checking the viability of their medications. The information and understanding of such factors are of epidemiological importance. The results will be beneficial in developing tools to support clinicians in the identification of factors related to HIV-Infected Patients. The results can be further use to shape communication and counseling strategies at the individual level before treatment initiation. Effective HAART initiation immediately after HIV exposure is necessary to suppress the increase of viral loads to induce potential ART benefits that accrue over time. Effective monitoring and modeling of disease biomarkers are essential to help inform methods that can be put in place to aimed to suppress viral loads for maximum ART benefits that can be accrued over time at an individual level. In this study, we have concentrated on the transformed normalized response data, which is the square root of CD4 count, that is continuous and conditional on the explanatory variables, and random effects have a normal distribution. Mixed models with random effects can also be applied to non-normal responses.

Chapter 3

Negative binomial mixed models for analyzing longitudinal CD4 count data

3.1 Introduction

A linear model consists of an outcome variable Y, which is assumed to be normally distributed, and several predictors $(X_1, X_2, \cdots, X_p)$. The simple linear model is given by

$$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2)$ and ϵ_i is independent for $i = 1, 2, \cdots, n$.

We can extend these linear model ideas to generalized linear models (GLM). Where we, too, have an outcome variable Y and predictor variable(s) X. The outcome Y can be continuous, dichotomous, count, ordinal, categorical, and so on as long as it comes from the exponential family. The exponential family of distribution incorporates numerous valuable distributions such as Poisson and Negative Binomial for count data; Binomial, Bernoulli, and Geometric for discrete data; Gamma, Normal, Inverse Gaussian, Exponential, and Beta for the study of continuous response data set. A distribution belongs to an exponential family of distribution if its probability density function (pdf) or probability mass function (pmf) can be expressed as

$$f(Y; \theta, \phi) = exp\left\{\frac{1}{a(\phi)}[\theta Y - b(\theta)] + c(Y, \phi)\right\} \tag{3.1}$$

where θ is the natural or canonical parameter, $a(\phi)$ is the scale parameter or dispersion and $c(Y, \phi)$ is some function of Y and ϕ. The mean, $\mu = E(Y) = b'(\theta)$, and the variance, $Var(Y) = \phi b''(\theta)$ (Dobson & Barnett, 2018).

Assume we have an outcome variable $\mathbf{Y} = [y_1, y_2, \cdots, y_n]$ that is expected to have the same distribution from an exponential family with $E[Y] = \mu$ and we have a set of parameters $\boldsymbol{\beta}$ with predictor variables $\mathbf{X}_i' = [x_{i1}, x_{i2}, \cdots, x_{ip}]$ that produces a linear predictor (η) such that $\eta = \mathbf{X}_i'\boldsymbol{\beta}$. The predictor variables and the outcome variable links to each other through the so-called "*link function*"($g(\cdot)$) such that $\eta = g(\mu)$. That is $g(\mu_i) = \eta_i = \sum_{i=1}^{p} x_{ij}\beta_j$. In general, there are three components to all generalized linear models (GLMs). These are the *random component* which identifies the response variable Y and assumes a probability distribution for it. The *systematic component* specifies the explanatory variables $(x_1, x_2, \cdots, x_p)$ used as predictors in the model, and the linear combination of the explanatory variables is called *linear predictor* (Casella & Berger, 2002). The *linear predictor* is given by $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p = \sum_{j=0}^{p} \beta_p x_p$. The *link function* describes the functional relationship between the *systematic component* and the expected value ($\mu = E(Y)$) of the *random component*. For linear models, this *link function* is simply the "*identity link*,"which means that Y itself is modeled, but there are some examples where that is not the case. Some of the examples of *link function* are summarized in (Fitzmaurice et al., 2008, 2012; Der & Everitt, 2012). *Identity link* which is the most straightforward possible link function that has the form $g(\mu) = \mu$. It is used for the continuous response ordinary regression model. *Log link* or log-linear model used for non-negative integer values such as count data. It has the form $g(\mu) = log(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$. *Log link* models the log of the mean (μ). This allows the mean to be non-linearly related to the predictors. The *logit link* or logit model is used when the outcome variable is dichotomous (usually 0 and 1). The fourth illustration of the link function is the *probit link*. The probit moreover demonstrates when μ is between 0 and 1, such as a probability $\phi^{-1}(\mu) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}$. The contrast between *logit* and *probit* link functions is regularly only seen in small samples, since the *probit link* assumes the normal distribution of the probability of an event, while the *logit link* assumes the logistic distribution (Menard, 2002; McCullagh & Nelder, 1989).

Note that the ordinary regression models for longitudinal data analysis lack as they fail to consider the reliance between observations over time. This means when data are measured repeatedly, like the CD4 count of an individual over time, the assumption of independence is no longer reasonable. The GLM is usually extended to generalized linear mixed models (GLMMs), with a subject-specific random effect added

in the linear predictor to capture the dependence (correlation).

Generalized linear mixed models (GLMMs) combine the above specification of generalized linear models with the linear mixed models discussed in Chapter 2. GLMMs include random effects into the linear predictor as an extension of generalized linear models. As an extension of the linear mixed model, GLMMs contain at least one fixed effect and at least one random effect. This permits the modeling of correlated, conceivably non-normally distributed data with a flexible settlement of covariates. This may overcome the modeling issue of overdispersion in the data and, at the same time, oblige the population heterogeneity. More particularly, let Y be the outcome variable whose conditional distribution given the random effects belongs to the exponential family, $x_1, \cdots, x_p$ be a set of p explanatory variables describing the fixed effects and let $u_1, \cdots, u_k$ be a set of q random effects. The linear predictor (η) of the model for the j^{th} observation given the random effects is expressed as

$$\eta_{ij} = g(E[Y_j|u_1, \cdots, u_q]) = \beta_0 + \sum_{i=1}^{p} \beta_i x_{ij} + \sum_{k=1}^{q} z_{kj} u_k, \qquad j = 1, \cdots, n \tag{3.2}$$

where β_0 is the intercept, β_i is the i^{th} fixed effect coefficient, x_{ij} is the i^{th} fixed effect explanatory variable on the j^{th} observation, z_{kj} is the binary indicator variable for the effect of the k^{th} random effect, u_k on the j^{th} observation, and $g(\cdot)$ is the link function relating the conditional mean of the response to the predictors.

Comparing the GLMMs specification to Section 2.4, the outcome variable is no longer required to be normally distributed. The relationship between the conditional mean of the response and the linear predictor is now on the *link* scale. This is an essential difference between these models and those where the response variable needs to be transformed before analysis (Gbur et al., 2012). In addition, when the observed data are counts, GLMMs are a viable strategy for dealing with such data. The most common cause of overdispersion with count data is assuming the wrong distribution for the model. The leading candidates for dealing with count data are the Poisson and the negative binomial distribution. Hurdle and Zero-Inflated regression models are utilized to handle the distribution of count data with excess zeros (Morel & Neerchal, 2012; Liu & Cela, 2008; Lambert, 1992; Mullahy, 1986). Since our motivating data set, illustrated in Section 3.7, does not have excess zeros thus, discussions of the Poisson and negative binomial distribution in the context of generalized linear mixed-effects models are the primary focus in the succeeding sections.

3.2 Marginal versus Conditional Models

The need to distinguish models according to the interpretation of their regression coefficients has led to the use of the terms *marginal models* and *conditional models* (Fitzmaurice et al., 2012). As with generalized linear mixed models (GLMMs), linear mixed models (LMMs) can also be formulated as *marginal* or *conditional* models (Diggle et al., 2002; Zeger & Liang, 1986; Molenberghs & Verbeke, 2000; Fitzmaurice et al., 2008, 2012). However, for GLMMs, the *marginal* versus *conditional* model is far more consequential because of the issue of overdispersion, which is unique to models for non-normally distributed data. Over-dispersion is discussed in Section 3.4.

The most crucial point of the marginal and conditional models for non-normally distributed GLMMs, unlike LMM, for normally distributed data is that they do not yield identical estimates. For non-normal data, they are not equivalent; they do not estimate the same thing (Diggle et al., 2002; Fitzmaurice et al., 2012). The estimated probabilities from the marginal model are often referred to as *population-averaged estimates* or marginal estimates. Predicted probabilities from the conditional model are variously called *subject-specific estimates*, *mixed-effects model estimates*, *random-effects model estimates*, or *conditional model estimates* (Diggle et al., 2002; Molenberghs & Verbeke, 2006; Hardin & Hilbe, 2003). The target of inference for the marginal model is the *population*, whereas the target of inference for the conditional model is the *individual* (Fitzmaurice et al., 2012). Given the different analytic focuses, the interpretation of the regression coefficients also differs markedly between the marginal and conditional perspectives in longitudinal data analysis. For the marginal models, the covariate regression coefficients represent an average effect on the linear predictor, which cannot directly be transformed into the population-averaged effect on the untransformed scale. For the conditional models, the regression coefficient of a covariate indicates the change in the transformed response variable (e.g., log, log odds) with a one-unit increase in the covariate within a subject (Liu, 2015). This section discusses marginal and conditional models, and the main focus is on the GLMMs.

3.2.1 Marginal Models

An alternative way of specifying mixed models is marginal models. As the name infers, they are characterized in terms of the marginal distribution of the observations. The term *marginal* in this setting demonstrates that the model for the mean response depends only on the fixed effects (covariates of interests) and not on any random effects (Fitzmaurice et al., 2012). That is, the term marginal is used to emphasize that the model for the mean response at each occasion does not rely on dependence

among observations. Such a marginal approach has tremendous appeal to many researchers of various disciplines who are concerned with the covariates' effects on the nonlinear response and to whom the impact of the random effects is not of direct interest (Liu, 2015). The random effects in marginal models are not modeled explicitly, but their impact on variation is embedded in the covariance structure of the model (Gbur et al., 2012). This is in contrast to conditional (mixed-effects) models, where the mean response depends not only on fixed-effects but also on a vector of random effects (Fitzmaurice et al., 2012). For this reason, marginal models are appropriate only when inference about the population average is the main focus (Diggle et al., 2002; Fitzmaurice et al., 2012).

The marginal LMMs and GLMMs share the same linear predictor, but the distribution and variance assumptions differ. The distribution applies exclusively to the response variable Y because there are no random effects on which to condition; all marginal GLMMs are defined on *quasi-likelihood*, not on exact probability distributions. All of the variance-covariance structure in the marginal GLMM uses a *working-correlation structure* (Diggle et al., 2002).

Marginal GLMMs can only be estimated using *quasi-likelihood* methods, whereas conditional models may use *quasi-likelihood* as well as an *integral approximation* (see Section 3.2.2). In general terms, marginal models do not require distributional assumptions for the observations, only a regression model for the mean response. That is, marginal models provide a unified method for analyzing diverse types of longitudinal responses by avoiding assumptions about the distribution of the vector of responses; the method relies solely on assumptions about how the mean response is related to the covariates. The avoidance of distributional assumptions leads to a method of estimation known as generalized estimating equations (GEE). Technically the term GEEs usually refers to the marginal models, whereas GLMMs refer to the conditional models (Fitzmaurice et al., 2012; Zeger & Liang, 1986; Liang & Zeger, 1986).

Strictly speaking, GEE refers to generalized linear models (GLMs) with no random effects in the linear predictor and all of the variance-covariance structures associated with the random factors embedded in the working correlation structure (Gbur et al., 2012). GEEs became very popular when GLMM computing software and computing technology, in general, were less developed (Stroup, 2012; Gbur et al., 2012). GEEs are still deeply entrenched in certain disciplines. Nonetheless, GEEs are useful if the conditional GLMMs are too complex to be computationally tractable or if the objectives of the study are best addressed by the marginal mean rather than the

conditional mean (Diggle et al., 2002; Stroup, 2012; Agresti, 2003). However, GEEs are associated with a lack of efficiency due to incomplete, occasionally, incorrect model specifications when the sample size is small or the regression model includes time-varying covariates (Liu, 2015; Fitzmaurice et al., 2012). Furthermore, based on the assumption of missing completely at random (MCAR), the GEE models cannot be applied effectively if missing data mechanisms are complex (Liang & Zeger, 1986; Liu, 2015; Fitzmaurice et al., 2012). Given the restrictions in the approach, GEEs have gradually become a much less applied methodology than GLMMs in the analysis of nonlinear longitudinal data (Liu, 2015). The use of GEE to estimate regression coefficients specified by marginal models has been studied extensively, for more details, see Liang & Zeger (1986), Zeger & Liang (1986), Fitzmaurice et al. (2012), Agresti (2003), McCullagh & Nelder (1989), Liu (2015), Diggle et al. (2002), Molenberghs & Verbeke (2000).

Marginal GLMMs

Let Y_{ij} denote the response variable for the i^{th} subject on the j^{th} measurement occasion. The response variable Y_{ij} can be continuous, binary, ordinal, or count. The nature of the response variable does have important implications for model specification; however, the notation does not distinguish among the diverse types of responses. Suppose that Y_{ij} is a count, and we wish to relate changes in the expected count (or expected rate) to the covariates. For this motivation, Fitzmaurice et al. (2012) discuss three illustrations of the marginal model for Y_{ij}:

- The mean of the j^{th} response, given $x_{i1}, \cdots, x_{in_i}$, depend only on X_{ij}.

$$E(Y_{ij}|x_{i1}, \cdots, x_{in_i}) = E(Y_{ij}|X_{ij}), \qquad i = 1, \cdots, N; \qquad j = 1, \cdots, n_i.$$

  This assumptions implies that given X_{ij}, there is no dependence of Y_{ij} on X_{ik} for $k \neq j$, and the mean of Y_{ij} is related to the covariates through a log link function, $log(\mu_{ij}) = \eta_{ij} = X'_{ij}\beta$, where β describes the change in the log of the population-average count per unit change in x.

- To account for the overdispersion, which is not prescribed by the Poisson model, one can assume the variance of each Y_{ij}, given the effects of the covariates, depends on the mean response, $Var(Y_{ij}|X_{ij}) = \phi\mu_{ij}$, where $\phi > 1$ is a time-variant scale parameter that needs to be estimated.

- The within-subject association among the vector of repeated responses is assumed to have an unstructured pairwise correlation pattern, $corr(Y_{ij}, Y_{ik}|X_{ij}, X_{ik}) = \alpha_{jk}$, where α is a correlation parameter, which is between 0 and 1. Here a

balanced longitudinal design is assumed, and the vector of parameters α represents the pairwise correlations among respondents.

3.2.2 Conditional Models

The name *conditional model* is derived from the fact that the distribution of the observations is specified conditionally on the random effects (Gbur et al., 2012). The conditional approach defines the probability distribution of the outcome variable as a function of the covariates and a parameter specific for each individual. The illustration of the conditional models is often based on the assumption that longitudinal data follow some particular stochastic distributions that reflect intraindividual dependence. Without the consideration of this distribution in regression modeling, the quality of parameter estimates, both point, and variance will be influenced by the correlation of repeated measurements for the same subject (Liu, 2015).

The conditional models give an effective approach for using fully specified probability functions to fit non-normal as well as normal distributed longitudinal data. These models are preferable, especially when the trajectory of non-normal distributed response outcomes is of primary interest (Liu, 2015). For longitudinal data, a prediction must combine the information of the estimated regression coefficients, the values of covariates, and the approximated random effects. This can be done under conditional models (Fitzmaurice et al., 2012). As each subject is assumed to have a unique random parameter, the random effect approximates are an integral component in the predictions. In contrast, the GEE models, where the variance-covariance matrix is specified as a nuisance parameter, cannot be used for nonlinear predictions except that all subjects potentially have the same value of the random effect parameter (Liu, 2015).

Conditional GLMMs

Consider longitudinal data with repeated measurements taken on the same individuals over time. Although the presence of different subjects is a known source of variation, the variability in the response due to the predictor variables $x_{i1}, ..., x_{in_i}$ is of greater interest. The interest lies not in the specific subjects that happened to be observed but rather in understanding the heterogeneity in the population of subjects and how it relates to the variability present in the data. The predictors $x_{i1}, ..., x_{in_i}$ have specific, fixed values of interest and are therefore known as *fixed effects*. The random effects, in contrast, are viewed as a random sample from a population of such effects. A model that incorporates both fixed and random effects is referred

to as a *mixed-effects model* (see Chapter 2). A generalized linear model that includes random effects is, therefore, referred to as a GLMM (Diggle et al., 2002; McCulloch et al., 2008).

In GLM, the conditional distribution of Y_{ij} given U_i follows a distribution from the exponential family with density $f(Y_{ij}|U_i;\beta)$ given U_i the repeated measurements, $y_{i1},...,y_{in_i}$, are independent; the $U_i \overset{iid}{\sim} f(U_i, G)$. Much like a GLM, a GLMM relates the mean of a response Y_{ij}, $i = 1,...,N$; $j = 1,...,n_i$, to a set of p predictors X_{ij} through a link function $g(\cdot)$. In addition to the fixed predictors X_{ij}, the linear component of a GLMM also includes q random effects U_i with q-variate density f_{U_i}. Conditional on U_i, one typically assume that the data are independent observations from a parametric distribution with density $f(Y_{ij}|U_i)$ and mean $E[Y_{ij}|U_i = u]$, and then models the conditional mean as: $\mu = E[Y_{ij}|U_i = u] = h(X'_{ij}\beta + Z'_{ij}u)$, where Z_{ij} is a q-vector of covariates associated with the random effects (Diggle et al., 2002). This model is hierarchical in structure and does not directly assume a marginal distribution for Y_{ij}. Rather, distributional assumptions are made for U_i and for $Y_{ij}|U_i$, and one must integrate over the random effects density f_{U_i} to obtain the marginal distribution for Y_{ij}. Therefore, the GLMM is a *conditional model* because the mean μ is conditioned on the random effects U_i.

Since the probability distribution of the response variable is specified conditionally on the random effects, parameters in GLMM can be fit using maximum likelihood estimation. However, for many choices of link function and random effects distribution, evaluating the marginal likelihood involves an analytically intractable integral. To overcome this issue, one could use numerical integration or maximize an approximation of the marginal likelihood instead of the true likelihood and thereby avoid the intractable integral. Alternative approximation methods are discussed in Section 3.3. Furthermore, the fixed effects parameters β have subject-specific interpretation, which means that each element of β provides information about the effect of the corresponding predictor on the response for a specific subject realization of the random effect. However, this interpretation does not always make sense because some predictors, such as indicators (e.g., gender), never change within a single subject.

3.3 Inference in Generalize Linear Mixed Models (GLMMs)

Frequently, the least-squares method has been used as the basis of estimation and statistical inference in linear models where the outcome variable is normally distributed. As an estimation method, least squares is a mathematical approach for minimizing the sum of squared errors that do not depend on the probability distri-

bution of the outcome variable. While they are appropriate for fixed effects models with normally distributed data, least squares do not generalize accurately to models with random effects, non-normal data, or both. This means that least-squares become gradually much less applicable as modeling complexity increases. Likelihood-based procedures are the typical strategy to solve these issues, which offers an alternative method that accommodates the probability distribution of the outcome variable into parameter estimation as well as inference.

Inference for mixed and generalized linear models is based on a likelihood method described in Subsection 2.4.5. Maximum likelihood (ML) or editions of ML, such as restricted maximum likelihood (REML), are standard methods of estimation for linear mixed models and generalized linear models. ML is also broadly used for fitting GLMMs and has the following properties: ML estimators are asymptotically normal, with standard errors available from second derivatives of the log-likelihood; Under broad conditions, ML estimators are asymptotically efficient; and Hypothesis testing can be carried out using likelihood ratio, score, or Wald tests (Stroup, 2012). Maximum likelihood estimation in GLMMs requires maximizing the marginal likelihood, which can be challenging because the likelihood to be maximized does not have a simple closed-form expression. It can be obtained either by averaging or integrating over the distribution of the random effects; maximizing that likelihood is challenging (Stroup, 2012). Therefore, the estimating equations cannot be written precisely. There are two techniques to doing this: *linearization*, specifically, the *quasi-likelihood method* and *integral approximation*, which are discussed in the later sections. Each technique can get very elaborate and computationally intense as the complexity of the GLMM increases. However, this simple illustration will serve our purpose: to have a conceptual sense of how the approximation works.

3.3.1 Quasi-Likelihood and Integral Approximation Methods

Quasi-Likelihood Method

In some statistical investigations, the distribution of the data is known. In others, we are less confident. With a little more experience with particular data, we would recognize that the variance increases with the mean, and we might have a tough concept as to how rapidly it increases. However, we are unlikely to know what distribution structure is correct or even possibly fit well. But not knowing the distribution makes it impossible to assemble a likelihood and hence use such techniques as maximum likelihood and likelihood ratio test. It would, therefore, be useful to have inferential methods that work as well or almost as well as maximum likelihood but without having to make specific distributional assumptions (McCulloch et al., 2008).

Wedderburn (1974) formalized this fundamental concept via *quasi-likelihood* theory to derive a likelihood-like extent whose development requires few assumptions, like properties of the *score function*.

Quasi-likelihood is defined as

$$Q_i = \int_{y_i}^{\mu_i} \frac{y_i - \mu_i}{a(\phi)\nu(\mu_i)} d\mu_i, \tag{3.3}$$

where $y_i = (y_1, \cdots, y_n)$, $\nu(\mu_i)$ corresponds to the form of the variance function, and $a(\phi)$ denotes the scalar parameter function, which is the unspecified constant of proportionality relating $Var(y_i)$ to $\nu(\mu_i)$ (McCulloch et al., 2008).

Letting $\nu(\mu_i) = \mu_i$ and $a(\phi) = 1$ in Equation (3.3) yields

$$= \int_{y_i}^{\mu_i} \frac{y_i - \mu_i}{\mu_i} d\mu_i = y_i log(\mu_i) - \mu_i,$$

which becomes the Poisson log-likelihood without $c(y, \phi) = -log(y!)$. Similarly, setting $\nu(\mu_i) = 1$ and $a(\phi) = \phi^2$ in Equation (3.3) yields

$$= \int_{y_i}^{\mu_i} \frac{y_i - \mu_i}{\phi^2} d\mu_i = \frac{y_i log(\mu_i) - \frac{\mu_i^2}{2}}{\phi^2},$$

which becomes the *quasi-likelihood* part of the Gaussian log-likelihood (Stroup, 2012).

Integral Approximation Methods: Laplace and Gauss-Hermite quadrature

The two commonly used integral approximation approaches for GLMMs are the Laplace approximation (McCulloch, 1997) and Gauss-Hermite quadrature (Pinheiro & Bates, 1995).

Laplace approximation

The idea of the Laplace approximation is to use a quadrature approximation at the point where the integrand takes its maximum. Its basic form is based on a second-order Taylor series expansion to obtain an analytical approximation of the integral, and the standard normal form of Laplace approximation takes as

$$= \int_{-\infty}^{+\infty} exp\{h(x)\}dx \cong \sqrt{2\pi} exp\{h(\tilde{x})\} \left(-\frac{\partial^2 h(x)}{\partial x^2}\bigg|_{x=\tilde{x}} \right)^{-\frac{1}{2}} \tag{3.4}$$

where $\tilde{x}$ denotes the value of x that maximizes $h(x)$, and by Taylor expansion, $h(x) = h(\tilde{x}) + h'(\tilde{x})(x - \tilde{x}) + \frac{h''(\tilde{x})(x-\tilde{x})^2}{2!} + \frac{h'''(\tilde{x})(x-\tilde{x})^3}{3!} + \cdots$, which yields an approximation to

$$\cong \sqrt{\frac{2\pi}{h''(\tilde{x})}} exp\{h(\tilde{x})\}$$

Gauss-Hermite quadrature

Gauss-Hermite quadrature is a method for approximating the integral of a function $f(\cdot)$ multiplied by another function having the shape of a normal density. The approximation is a finite weighted sum that evaluated the function at certain points (Stroup, 2012). For a set of nodes $x_1, \cdots, x_n$ and weights $w_1, \ldots, w_n$, the Gauss-Hermite quadrature approximation has the form

$$= \int_{-\infty}^{+\infty} f(x) e^{-x^2} dx \cong \sum_{k=1}^{n} w_k f(x_k), \tag{3.5}$$

the *weights* $\{w_k\}$ and *quadrature points* $\{x_k\}$ are tabulated in standard reference books of mathematical tables such as Zwillinger (2002), and Abramowitz & Stegun (1948), or can be computed with a formula provided by Golub & Welsch (1969). The approximation with Gauss-Hermite quadrature improves as k, the number of quadrature points increases. However, increasing k also increases the procedure's computational burden, to the point where it becomes prohibitive (Pinheiro & Bates, 1995; Stroup, 2012). Although statistical software packages such as SAS PROC GLIMMIX procedure are *adaptive*, that is, they have data-driven decision rules to select a nominally optimal number of quadrature points. In some situations, the complexity of the model makes the adaptive procedure itself computationally restrictive (Stroup, 2012).

A detailed discussion of *quasi-likelihood*, Laplace approximation, Gauss-Hermite quadrature, and other alternative approximation methods of estimation for GLMMs can be found in numerous literature (Wedderburn, 1974; Stroup, 2012; McCulloch, 1997; Fitzmaurice et al., 2012; Jiang, 2007; McCulloch et al., 2008; Breslow & Clayton, 1993; Demidenko, 2013; McCullagh & Nelder, 1989; Pinheiro & Bates, 1995). Searle and McCulloch, 2001. Statistical software packages: SAS PROC GLIMMIX Procedure, and R such as the *lme4, MASS, glmmPQL, glmmADMB*, and *glmmML* packages can fit these approximation methods for various GLMMs.

3.4 Issues of Overdispersion in GLMMs

The term overdispersion would imply more variability shown by the data than would be assumed under a given statistical model. For instance, if a response variable, Y, is count (positive integer) and assumed to have a Poisson distribution, then, in theory, we implicitly assume that $E(Y) = \lambda = Var(Y)$, which is also known as *equidispersion*. However, if the sample variance exceeds the sample mean, then the data are said to be overdispersed; that is, the observed variance is implausibly large for the Poisson assumption to be correct. This indicates that not all processes that give rise to count data can be modeled as Poisson. In some cases, the total count is bounded, in which case a binomial distribution should probably be used. In other cases, the counts may be adequately large that a normal approximation is advocated so that a normal linear model might be used.

Overdispersion is an issue that should not be disregarded in the analysis. The essential and most serious consequences of failing to account for overdispersion are underestimating of standard errors and inflate test statistics; consequently, excessive type I error rate and inadequate confidence interval coverage (Stroup, 2012). We illustrated this in Table 3.6, uncorrected analysis of overdispersed data results in underestimated standard errors, leading to biased estimates and inflated statistics. It is necessary to check for overdispersion when fitting a GLM or a GLMM to guarantee that inferences derived from the fitted model are precise (Morel & Neerchal, 2012; Molenberghs et al., 2007).

Overdispersion is an implication that the fitted model is incorrect, and adjustments are required. The fitted model may be inaccurate by the improper choice of any of the three components in GLMMs: the linear predictor, the distribution of the observed data, or the link. Also, for GLMMs, overdispersion is associated with the variance and covariance assumptions for the random effects (Stroup, 2012). For this reason, action needs to be taken to avoid the unwanted outcomes outlined above. The two most commonly used approaches in GLMMs are: adjusting the standard errors and test statistics by incorporating an adjustment for overdispersion in the assumed model or consider a different probability distribution for the observed data that more reasonably approximate the method by which overdispersion emerge (McCullagh & Nelder, 1989). Because the second strategy of assuming a different distribution is a reasonable and suggested methodology, we illustrated this in Table 3.6 in which the negative binomial distribution (see Section 3.6) substitutes the Poisson distribution (see Section 3.5) as the conditional distribution of the count outcome variable.

Detecting diagnostics such as *residual plots* and *Pearson* χ^2/df fit statistics can be computed from the data to assess overdispersion. In ordinary least squares (OLS), *residual* refers to the difference between the observed and its fitted value. However, in generalized linear mixed models, *residuals* are scaled in two different ways: on the model scale and the data scale. The *Pearson* and *Studentized residual* using the estimated variance of the conditional distribution and residual, respectively, on each scale, summarized in Table 3.1 below.

Table 3.1: Summary of residuals in GLMMs

Residual	Model scale	Data scale
Pearson	$\frac{\boldsymbol{y^*-\hat{\eta}}}{\sqrt{\hat{Var}(\boldsymbol{y^*}\|\boldsymbol{b})}}$	$\frac{\boldsymbol{y-h(\hat{\eta})}}{\sqrt{\hat{Var}(\boldsymbol{y}\|\boldsymbol{b})}}$
Studentized	$\frac{\boldsymbol{y^*-\hat{\eta}}}{\sqrt{\hat{Var}(\boldsymbol{y^*-\hat{\eta}})}}$	$\frac{\boldsymbol{y-h(\hat{\eta})}}{\sqrt{\hat{Var}(\boldsymbol{y-\hat{\mu}})}}$

Where $\boldsymbol{y^*}$ is the *pseudo-variate*, $\boldsymbol{h(\cdot)}$ is the inverse link function, $\boldsymbol{\hat{\eta}}$ is the estimated linear predictor, and $\boldsymbol{\hat{\mu}}$ is the estimated mean (Stroup, 2012).

3.5 Poisson Regression Model in the Context of GLMMs

Poisson Regression Model is one of the special case of generalized linear models, which share the following features: The mean response $\mu_i = E(Y_i)$, is assumed to be related to a vector of covariates, $\boldsymbol{x}$, through $h(\mu_i) = \boldsymbol{x_i'\beta} = log(\mu_i)$; the function, $h(\cdot)$, is called the *link function*. The variance of $\boldsymbol{Y_i}$ ia a specific function of its mean, μ_i, namely, $Var(Y_i) = \nu_i = \phi\nu(\mu_i)$, the function $\nu(\cdot)$ is known and referred to as the variance function; ϕ is the scaling factor, which is a known constant for some members of the GLM family, whereas in others it is an additional parameter to be estimated. Finally, the Poisson regression model is a member of the exponential family of distribution, with a likelihood of the form expressed in Equation (3.1), can be shown as follows: Note that: λ^Y=$exp\{Yln\lambda\}$ and $\frac{1}{Y_i!} = exp\{-lnY_i!\}$. Therefore,

$$f(Y_i, \lambda_i) = exp\{\frac{Y_i ln\lambda_i - \lambda_i}{1} - lnY_i!\}$$

where $\phi = 1$, the canonical parameter $\theta = ln\lambda_i$, $b(\theta) = \lambda$, $c(Y, \phi) = -lnY_i!$.

Let Y_i be a response variable and $x_i = (x_{i1}, ..., x_{ip})'$ be a $p \times 1$ vector of covariates for the i^{th} individual then the pdf of the Poisson regression model with parameter λ_i is

given by

$$f(Y_i, \lambda_i) = \frac{e^{-\lambda_i}\lambda_i^{Y_i}}{Y_i!} \tag{3.6}$$

where $\lambda_i = exp(\boldsymbol{x_i'\beta})$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)'$ is a $p \times 1$-dimensional vector of unknown parameters corresponding to $\boldsymbol{x_i}$.

In the statistics literature the model comprising in Equation (3.6) is also called a *log-linear* model since the logarithm of the conditional mean, $E[Y_i|x_i] = \lambda_i = exp(\boldsymbol{x_i'\beta})$, is linear in the parameters: $\ln E[Y_i|x_i] = ln\lambda_i = \boldsymbol{x_i'\beta}$.

The Poisson regression model figures prominently in the modeling of count data. Count data comes from counting events of interest in an experimental unit, especially increasingly common in biostatistical science. For instance, the number of COVID-19 infected patients recorded during the coronavirus prevention programme. Counts are non-negative integer, often right-skewed, with a Poisson or negative binomial distribution. The Poisson regression is a commonly-used statistical model for n responses $y_1, \cdots, y_n$ that take count values. Each y_i is modeled as an independent Poisson (λ_i) r.v. and distributed as $y_i \overset{iid}{\sim}$ Poisson (λ_i), where the parameter λ_i controls the count rate in the i^{th} time. Thus, a model for the Poisson rate parameter λ_i is given by

$$ln\lambda_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \sum_{j=1}^{p} \beta_j x_{ij}$$

or equivalently,

$$\lambda_i = exp\{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\} = exp\{\sum_{j=1}^{p} \beta_j x_{ij}\} \tag{3.7}$$

where $x_{i1}, \cdots, x_{ip}$ are set of p covariates, and $\beta = (\beta_0, \cdots, \beta_j)$ are the regression coefficients.

Since $y_i \overset{iid}{\sim}$ Poisson (λ_i), as a consequence, the likelihood function is equal to the product of their pdf:

$$L(y_1, \cdots, y_n|\lambda_i) = \prod_{i=1}^{n} f(Y_i, \lambda_i) = \prod_{i=1}^{n} \frac{e^{-\lambda_i}\lambda_i^{Y_i}}{Y_i!} = \lambda_i^{\sum_{i=1}^{n} Y_i} e^{-n\lambda_i} \prod_{i=1}^{n} \frac{1}{Y_i!}$$

The log-likelihood function can be derived by taking the natural logarithm of the likelihood function:

$$\ell(y_1, \cdots, y_n|\lambda_i) = ln\left(\lambda_i^{\sum_{i=1}^n Y_i} e^{-n\lambda_i} \prod_{i=1}^n \frac{1}{Y_i!}\right) = ln(\lambda_i)\sum_{i=1}^n Y_i - n\lambda_i - \sum_{i=1}^n lnY_i!$$

$$= \sum_{i=1}^n [Y_i ln(\lambda_i) - \lambda_i - lnY_i!] \tag{3.8}$$

where λ_i is defined interms of $\beta_0, \cdots, \beta_p$ and the covariates $x_{i1}, \cdots, x_{ip}$ in Equation (3.7). Setting $x_{i0} \equiv 1$ for all i, the log-likelihood function can be expressed as

$$\sum_{j=1}^p [Y_i(\beta_j x_{ij}) - exp(\beta_j x_{ij}) - lnY_i!] \tag{3.9}$$

The maximum likelihood estimator (MLE), which is the standard estimator of the Poisson rate (λ_i) is the solution of the following maximization problem:

$$\hat{\lambda_i} = \underset{\lambda}{\operatorname{argmax}}\, \ell(y_1, \cdots, y_n|\lambda_i)$$

The first order condition for a maximum is $\frac{\partial}{\partial\lambda}\ell(y_1, \cdots, y_n|\lambda_i) = 0$. The first derivative of the log-likelihood with respect to the parameter λ_i is

$$\frac{\partial}{\partial\lambda}\ell(y_1, \cdots, y_n|\lambda_i) = \frac{\partial}{\partial\lambda}(ln\lambda_i \sum_{i=1}^n Y_i - n\lambda_i - \sum_{i=1}^n lnY_i!) = \frac{1}{\lambda_i}\sum_{i=1}^n Y_i - n$$

Impose that the first derivative equal to zero, and we get the MLE of λ for the i^{th} observation, i.e. $\hat{\lambda_i} = \frac{1}{n}\sum_{i=1}^n Y_i$.

The first and second partial derivatives, respectively, of the $logL$ expressed in Equation (3.9) with respect to unknown parameter β are given by

$$\frac{\partial\ell}{\partial\beta} = \sum_i^n (Y_i - \lambda_i)\boldsymbol{x_i} \tag{3.10}$$

and

$$\frac{\partial^2\ell}{\partial\beta\partial\beta'} = -\sum_i^n \lambda_i \boldsymbol{x_i}\boldsymbol{x_i'} \tag{3.11}$$

3.5.1 Poisson mixed-effects model for longitudinal count data

The conventional Poisson regression model for count response, discussed in the above section, assume statistical independence of observations. However, in many cases, the frequency data are longitudinal, and the assumption of independence is no longer reasonable.

Suppose that there are $N = \sum_{i=1}^{n} n_i$ non-negative counts y_{ij} for $i = 1, ..., N$ subjects and $j = 1, ..., n_i$ observations for subject i and a p-dimensional unknown parameter vector, β, associated with a covariate vector $x_{ij} = (x_{ij1}, ..., x_{ijp})'$. For simplicity, consider a model with a single random effect ν_i, and assume that ν_i is normally distributed with mean 0 and variance σ^2 and independent of the covariate vector x_{ij}. Thus, the Poisson mixed-effects model conditional on the density function of the n_i individual responses for subject i is written as

$$f(\boldsymbol{y_i}|\theta) = \prod_{j=1}^{n_i} f(Y_{ij}; \lambda_i) = \prod_{j} \frac{exp(-\lambda_{ij})\lambda_{ij}^{Y_{ij}}}{Y_{ij}!} \tag{3.12}$$

where $\theta_i = Y_i/\sigma$ such that $\theta_i \sim N(0, 1)$, and $\lambda_{ij} = exp(\boldsymbol{x'_{ij}\beta} + \nu_i) = exp(\boldsymbol{x'_{ij}\beta} + \sigma\theta_i)$. The log-likelihood function corresponding to the above equation becomes

$$logL(\boldsymbol{y_i}|\theta) = \sum_{j=1}^{n_i} \left[Y_{ij}(\boldsymbol{x'_{ij}\beta} + \sigma\theta_i) - exp(\boldsymbol{x'_{ij}\beta} + \sigma\theta_i) - log(Y_{ij}!)\right] \tag{3.13}$$

Pertinent references about the description of the Poisson mixed-effects model include Breslow (1984), Lawless (1987), Dean et al. (1989), Stukel (1993), Thall (1988), and Liu (2015).

3.6 Negative Binomial Regression Model in the Context of GLMMs

The basic regression model to analyze count data is the Poisson model. However, the Poisson regression model is limited because it forces the conditional mean of the response to equal the conditional variance. This assumption is often violated in real-life data. Real-life count data usually feature overdispersion relative to the Poisson model. As previously discussed (Section 3.4), accounting for overdispersion when modeling count data is essential. Failure to cope with this feature of the data can lead to biased parameter estimates and thus false conclusions.

A commonly used model for overdispersed count data, where the variance exceeds

the mean, is the negative binomial model. As a generalization of Poisson regression, negative binomial regression loosens the restrictive assumption, which is the variance and the mean made by the Poisson model is equal by including a dispersion parameter to accommodate the unobserved heterogeneity in the count data. Here, we assume that given a rate μ_i, the Y_{ij} are independent Poisson variates with mean and variance equal to μ_i. The overdispersion arises because the $\mu_i's$ are assumed to vary across subjects according to a gamma distribution with mean μ and variance $\phi\mu^2$ which exceeds the Poisson variance when $\phi > 0$.

Like most regression models, the negative binomial regression is based on an underlying probability distribution function (*pdf*). For instance, the normal linear regression model is derived from the Gaussian (normal) *pdf*, and the Poisson regression is derived from the Poisson *pdf*. However, the conventional negative binomial model, which is commonly symbolized as NB2 or *quadratic* negative binomial based on the exponent in its second term (Cameron & Trivedi, 2013; Hilbe, 2014), is derived from a Poison-gamma mixture distribution. But such a mixture of distributions is only one of the ways in which the negative binomial *pdf* can be obtained. As we would in referring to the Poisson regression model or logistic regression model, the negative binomial model is not based on a single model (one derivation) (Hilbe, 2014). The negative binomial distribution can be mathematically derived from the binomial, Poisson inverse Gaussian, as well as from the geometric distributions. There are also more derivations of the negative binomial model. Some separate types of derivations for the negative binomial model were discussed here (Cameron & Trivedi, 1986, 2013; Boswell, 1970; Shoukri et al., 2004; Hilbe, 2011, 2014; Demidenko, 2013).

The standard negative binomial (NB2) model, predominantly as a Poisson-gamma mixture model with a mean of μ and a variance of $\mu + \alpha\mu^2$, is nearly always used to estimate parameters of overdispersed count data (Hilbe, 2014). The derivation of the Poisson-gamma mixture model can be addressed as follows: parameterization of the negative binomial regression as summarized by Demidenko (2013), is frequently expressed in terms of the mean λ, dispersion parameter θ, and a non-negative integer y. Let Y takes discrete values with the conditional Poisson distribution, $P_r(Y = y|\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}$, where $\lambda > 0$, $\lambda \sim Gamma(\alpha, \theta)$ then the *pdf* of a two-parameter, α, and θ, Gamma distribution is given by:

$$f(\lambda; \alpha, \theta) = \frac{\lambda^{\alpha-1} e^{-\frac{\lambda}{\theta}}}{\theta^{\alpha}\Gamma(\alpha)}, \qquad \lambda > 0, \quad \alpha > 0, \quad \theta > 0 \tag{3.14}$$

Thus, the negative binomial (Poisson-Gamma) model (joint *pdf* of Y and λ) can be defined as

$$f(y,\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!}\frac{\lambda^{\alpha-1}e^{-\frac{\lambda}{\theta}}}{\theta^{\alpha}\Gamma(\alpha)} \tag{3.15}$$

The marginal distribution of Y can be obtained by integrating out λ:

$$f(y) = \int_0^{\infty} \frac{e^{-\lambda}\lambda^{y}}{y!}\frac{\lambda^{\alpha-1}e^{-\frac{\lambda}{\theta}}}{\theta^{\alpha}\Gamma(\alpha)}\partial\lambda$$

$$= \frac{1}{\theta^{\alpha}y!\Gamma(\alpha)}\int_0^{\infty} e^{-\lambda}\lambda^{y+\alpha-1}e^{-\frac{\lambda}{\theta}}\partial\lambda$$

Let $\frac{\lambda}{\theta} = u$, $\frac{\partial\lambda}{\theta} = \partial u$, $\partial\lambda = \theta\partial u$. Thus, the above equation can be expressed as

$$= \frac{1}{\theta^{\alpha}y!\Gamma(\alpha)}\int_0^{\infty} e^{-\theta u}(\theta u)^{y+\alpha-1}e^{-u}\partial u$$

$$= \frac{\theta^{y}\theta^{\alpha}}{\theta^{\alpha}y!\Gamma(\alpha)}\int_0^{\infty} e^{-\theta u}u^{y+\alpha-1}e^{-u}\partial u = \frac{\theta^{y}}{y!\Gamma(\alpha)}\int_0^{\infty} e^{(-1+\theta)u}u^{y+\alpha-1}\partial u$$

Let $(1+\theta)u = z$, $(1+\theta)\partial u = \partial z$, $\partial u = \frac{\partial z}{(1+\theta)}$. Thus,

$$= \frac{\theta^{y}}{y!\Gamma(\alpha)}\int_0^{\infty} e^{-z}\left(\frac{z}{1+\theta}\right)^{y+\alpha-1}\frac{\partial z}{(1+\theta)} = \frac{\theta^{y}}{y!\Gamma(\alpha)}\int_0^{\infty}\frac{z^{y+\alpha-1}e^{-z}}{(1+\theta)^{y+\alpha}}\partial z$$

$$= \frac{\theta^{y}}{(1+\theta)^{y+\alpha}y!}\Gamma(\alpha)\int_0^{\infty} z^{y+\alpha-1}e^{-z}\partial z$$

where Γ is the gamma function which has the formula $\Gamma(\alpha) = \int_0^{\infty} z^{\alpha-1}e^{-z}\partial z$, for any positive real number α. Thus, the above equation becomes

$$\frac{\theta^{y}\Gamma(\alpha+y)}{(1+\theta)^{y+\alpha}y!\Gamma(\alpha)} = \frac{(\alpha+y-1)!\theta^{y}}{y!(\alpha-1)!(1+\theta)^{y+\alpha}} \tag{3.16}$$

which is a negative binomial density. It has also been defined in the literature as:

$$= \binom{\alpha+y-1}{y}\left(\frac{\theta}{1+\theta}\right)^{y}\left(\frac{1}{1+\theta}\right)^{\alpha} = \frac{\Gamma(\alpha+y)}{y!\Gamma(\alpha)}\left(\frac{\theta}{1+\theta}\right)^{y}\left(\frac{1}{1+\theta}\right)^{\alpha}, \tag{3.17}$$

where the binomial coefficient is computed as $\binom{\alpha+y-1}{y} = \frac{(\alpha+y-1)(\alpha+y-2)\cdots\alpha}{y!} = \frac{(\alpha+y-1)!}{y!(\alpha-1)!}$. Note that for a positive integer α, we have $\Gamma(\alpha) = (\alpha-1)!$.

For negative binomial distribution, $E(y) = \alpha\theta$, and $var(y) = \alpha\theta(1+\theta)$. For Poisson distribution, the mean and variance are equal, but for negative binomial, the variance is higher than the mean by $\alpha\theta^2$. By applying some calculus, one can show that the Poisson distribution is a special case of the negative binomial distribution when

$\alpha \to \infty$ and $\theta \to 0$, such that the product, $\alpha\theta = \lambda$, is kept constant. The parameter $a = \frac{1}{\alpha}$ is associated with the "extra-Poisson"variation, or overdispersion, because $var(y) = \lambda + a\lambda^2$, which is quadratic in the mean that is why the negative binomial model is referred to as the NB2 model as we mentioned previously. This interpretation justifies a (λ, a) parameterization of the negative binomial distribution as

$$P_r(Y = y; \lambda, a) = \binom{y + \frac{1}{a} - 1}{y} \left(\frac{a\lambda}{1 + a\lambda}\right)^y \left(\frac{1}{1 + a\lambda}\right)^{\frac{1}{a}}, \tag{3.18}$$

where $E[y] = \lambda$ and $var[y] = \lambda + a\lambda^2$, and $a = 0$ leads to Poisson distribution. This latest parameterization is convenient for specifying the negative binomial regression and for testing overdispersion as $H_0 : a = 0$ (Lawless, 1987).

The likelihood function of Equation (3.17) is proportional to

$$L(\theta, \alpha) = \prod_{i=1}^{n} \frac{\Gamma(\alpha + y_i)}{y_i!\Gamma(\alpha)} \left(\frac{\theta}{1 + \theta}\right)^{y_i} \left(\frac{1}{1 + \theta}\right)^{\alpha} \tag{3.19}$$

Lawless (1987) notes that for any $c > 0$, $\Gamma(y + c)/\Gamma(c) = c(c + 1) \times \cdots \times (c + y - 1)$ for integer-valued $y \geq 1$, thus, $\frac{\Gamma(\alpha+y)}{\Gamma(\alpha)} = \alpha(1 + \alpha) \times \cdots \times (y - 1 + \alpha)$. Hence, $log\left\{\frac{\Gamma(\alpha+y)}{\Gamma(\alpha)}\right\} = \sum_{i=1}^{y-1} log(1 + \alpha)$. This produces $logL(\theta, \alpha)$ as follows

$$= \sum_{i=1}^{n} \left(\sum_{i=1}^{y_i - 1} \log(1 + \alpha) - \log y_i! + y_i \log\theta - y_i \log(1 + \theta) + \alpha log1 - \alpha log(1 + \theta) \right)$$

$$\ell(\theta, \alpha) = \sum_{i=1}^{n} \left(\sum_{i=1}^{y_i - 1} \log(1 + \alpha) - \log y_i! + y_i \log\theta - (y_i + \alpha) \log(1 + \theta) \right)$$

Therefore, applying the Poisson theorem with Gamma distribution leads to the negative binomial distribution. Furthermore, detailed discussions of estimating methods and characteristics of the negative binomial regression model presented in numerous literature (Lord et al., 2012; Demidenko, 2013; Guide, 2008; Hilbe, 2011; Liu & Cela, 2008; Lawless, 1987).

When repeated counts are measured on the same individual over time, the assumption of independence is no longer reasonable; instead, they are correlated. Subject-specific random effects can be added into the linear predictor to modeling such dependence. Let y_{ij} be the values of a count variable (non-negative integer value) for subject i at time point j. The count is assumed to be drawn from a Poisson distribution with errors assumed to have a normal distribution, $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$. Then, the

Poisson mixed-effects model that specifies the expected number of counts is written as

$$\log(\mu_{ij}) = \boldsymbol{x'_{ij}}\boldsymbol{\beta} + \boldsymbol{z'_{ij}}\boldsymbol{b_i} + \varepsilon_{ij},$$

where $\boldsymbol{x'_{ij}}$ is the variable of interest, $\boldsymbol{\beta}$ is the vector of fixed effects (population-level effects), including an intercept β_0, $\boldsymbol{b}_i$ is the vector of random effects (subject-level effects) for the sample variables $\boldsymbol{z}_{ij}$, and ε_{ij} is the random errors (Fitzmaurice et al., 2012; Liu, 2015). Given the Poisson process for the count y_{ij}, the probability that $y_{ij} = y$, conditionally on the random effects $\boldsymbol{b}_i$, is given by

$$\begin{aligned} P(y_{ij} = y|\boldsymbol{b}_i, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}) =& \frac{e^{-\mu_{ij}}\mu_{ij}^{y}}{y!} = \frac{1}{y!}e^{-\exp(\boldsymbol{x'_{ij}}\boldsymbol{\beta}+\boldsymbol{z'_{ij}}\boldsymbol{b_i})}\exp(\boldsymbol{x'_{ij}}\boldsymbol{\beta} + \boldsymbol{z'_{ij}}\boldsymbol{b_i})^y \\ =& \frac{1}{y!}\exp[(\boldsymbol{x'_{ij}}\boldsymbol{\beta} + \boldsymbol{z'_{ij}}\boldsymbol{b_i})^y - \exp(\boldsymbol{x'_{ij}}\boldsymbol{\beta} + \boldsymbol{z'_{ij}}\boldsymbol{b_i})], y = 0, 1, 2, ... \end{aligned}$$

This addition also can be applied to the NBMM that allows over-dispersion by assuming a gamma distribution for the errors; instead of a normal distribution. Suppose that $\boldsymbol{x_{ij}}$ and $\boldsymbol{z_{ij}}$ are known vectors of covariates associated with count data $y_{ij}, i = 1, ..., n$ and $j = 1, ..., n_i$, conditional on a q-dimensional vector of subject-specific random effects, $\boldsymbol{b_i}$, the counts of y_{ij}, with the assumption of gamma errors, has a negative binomial distribution, $y_{ij}|\boldsymbol{b_i} \sim NB(\mu_{ij}, \mu_{ij} + \theta\mu_{ij}^2)$, with $\mu_{ij} = E(y_{ij}|\boldsymbol{b_i}) = \exp\{\boldsymbol{x'_{ij}}\boldsymbol{\beta} + \boldsymbol{z'_{ij}}\boldsymbol{b_i}\}$. This indicates that the mean parameters μ_{ij} of the negative binomial mixed-effects models are also related to the predictor variables $\boldsymbol{x_{ij}}$, and the sample variables $\boldsymbol{z_{ij}}$ through the logarithm link function: $\log(\mu_{ij}) = \boldsymbol{x'_{ij}}\boldsymbol{\beta}+\boldsymbol{z'_{ij}}\boldsymbol{b_i}+\varepsilon_{ij}$, which shows that the model for the conditional mean of the NBMM is similar to that of PMM. However, the conditional variance of y_{ij} for NBMM is $Var(y_{ij}|\boldsymbol{b_i}) = \mu_{ij} + \theta\mu_{ij}^2$, which is greater than the conditional mean of PMM by $\theta\mu_{ij}^2$, specifically, because a gamma distribution is assumed for the exponentiated errors, $\exp(\varepsilon_{ij})$, with a mean of 1 and variance θ (Fitzmaurice et al., 2012; Demidenko, 2013). Random effects are used to demonstrate multiple assets of variations and subject-specific effects. As a result, they avoid biased inference on the fixed effects. The random effects are assumed to have a multivariate normal distribution:

$$\boldsymbol{b_i} \sim N(0, \Psi),$$

where Ψ is a positive-definite variance-covariance matrix that accounts for the correlation of the random effects (Zhang et al., 2017, 2018).

3.6.1 Parameter Estimation and Model Selection in GLMMs

Several methods are available to estimate the parameters ($\beta_i's$ and $b_i's$) in GLMMs. We have listed a few of these methods herein: marginal quasi-likelihood (MQL), penalized (predictive) quasi-likelihood (PQL), the Laplace approximation, the Gauss-Hermite quadrative, and the Markov Chain Monte Carlo (MCMC) method (Guide, 2008; Gill & Torres, 2019). Our preference is for the Laplace approximation due to the fewer limitations and regularly equal or better results than the Adaptive quadrature (method=quad). Additionally, it is accurate, fast, and gives us the plausibility to urge likelihood and information criteria (Shoukri, 2018; Schabenberger & Gotway, 2017; Jiang, 2007). However, R-side random effects are not supported for method=laplace or method=quad in the Proc Glimmix statement. Instead, Proc Glimmix uses a random statement and the *residual option* to model repeated (R-side) effects.

Several software packages make it conceivable to perform GLMMs in R, such as *MASS* package with the function *glmPQL*, *lme4* package with the function *glmer*, and *MCMglmm* package with the function *MCMC*. We could also use programs such as SAS Proc GLIMMIX (method= *Laplace*), *WinBUGS Bayesian inference (MCMC)*, and nowadays, we can also use *SPSS* to do GLMMs. If we use either the *lme4* package with the function *glmer* or SAS Proc GLIMMIX (Laplace), we will get the Laplace approximation, which has an advantage that it gives *likelihood* and *IC*. The advantage of having those accessible is that we can compare several methods, observe what the impact is, including or expelling a fixed or a random effect by comparing the AIC, BIC, CAIC, or QIC, and the likelihoods depending on the adopted modeling strategy. The parameter estimates based on the mixed-effects negative binomial model are not exceptionally different from those based on the mixed-effects Poisson model. However, the Poisson model underestimates the standard errors when overdispersion is present, driving to improper inference. A straightforward way to select between these two models is to compare them based on a few criteria, such as AIC and BIC. Where for the ICs, a lower value better means that the model fits way better. We may moreover compare models utilizing $-2loglikelihood$ and the *likelihood ratio test*, and then if a model is significantly lower, it implies that it is the best model. Also, the regression parameters in GLMMs have somewhat different interpretations than the regression parameters in the conventional marginal models. In GLMMs, the regression coefficients have subject-specific interpretations. Especially, they represent the impact of variables on a particular subject's mean response. Specifically, the $\beta's$ are interpreted in terms of the effects of within-subject changes in explanatory variables on changes in an individual's transformed mean response while holding the remaining covariates constant. Accordingly, β_p can be interpreted as the change in

an individual's log of response for a unit increase in X_{pij} while holding other fixed variables constant for that individual. Since the components of the fixed effects, β, have interpretations that depend on holding b_i, the i^{th} individual's random effects, fixed, they are regularly referred to as subject-specific regression coefficients. Thus, generalized linear mixed-effects models are most valuable when the main scientific objective is to make inferences about individuals rather than the population average; the population averages are the targets of inference in marginal models (Fitzmaurice et al., 2012).

3.7 Data example: CAPRISA 002 Acute Infection Study data

In this section, we illustrate the performance of the methods discussed in the above sections on the CAPRISA 002 Acute Infection Study. The data is an ongoing prospective cohort study conducted on HIV-infected women at the Doris Duke Medical Research Institute (DDMRI) at the Nelson R Mandela School of Medicine of the University of KwaZulu-Natal in Durban, South Africa. Between August 2004 and May 2005, CAPRISA initiated a cohort study enrolling high-risk HIV-negative women to follow up. In the case of the data used in this paper as part of an ongoing study, women infected with HIV are enrolled in the study early, followed intensely, and monitored closely to study disease progression and CD4 count/viral load evolution. One can refer to studies by (Van Loggerenberg et al., 2008; Mlisana et al., 2014) for details on the design, development, and procedures of the study population.

Table 3.2 shows the summary of CD4 count and its associated selected covariates in the CAPRISA 002 Acute Infection Study. The dataset included 235 subjects (7129 observations consist of a minimum of two and a maximum of sixty-one observations per subject). p-values demonstrated in Table 3.2 are obtained from the Chi-square test. At a significance level of $\alpha = 5\%$, the univariate cross-tabulation analysis uncovers that the patient's baseline BMI, baseline viral load, number of sex partners, age, ART initiation, and level of education are significantly associated with the patient's CD4 count. Table 3.2 demonstrates that there is a high prevalence of CD4 count above 500 $cells/mm^3$ among patients having normal weight and overweight status, which are 38.32 and 9.36%, respectively (p-value<0.0001). Out of 7129 observations, patients having an undetectable viral load at baseline indicates no sign of a CD4 count below 500 $cells/mm^3$ throughout the study.

Table 3.2: Distribution of CD4 count and associated selected covariates with percent missing

		CD4 count N(%)				
Covariates	**Level**	<200	200-500	>500	**p-value**	**% Missing**
	Underweight	2(0.03)	219(3.12)	254(3.62)	<0.0001	0.0
Baseline	Normal weight	114(1.62)	2305(32.84)	2690(38.32)		
BMI Category	Overweight	18(0.26)	512(7.29)	657(9.36)		
	Obese	0	17(0.24)	231(3.29)		
	Undetected	0	0	16(0.23)	<0.0001	0.0
Baseline	Low	20(0.28)	791(11.27)	1532(21.83)		
Viral Load	Medium	45(0.64)	1209(17.22)	1497(21.23)		
	High	69(0.98)	1053(15)	787(11.21)		
Number of	No Partner	29(0.41)	565(8.05)	579(8.25)	<0.0001	0.0
sexual partners	Stable Partner	85(1.21)	2274(32.4)	3078(43.85)		
	Many Partner	20(0.28)	214(3.05)	175(2.49)		
	<20	1(0.01)	130(1.82)	121(1.72)	<0.0001	0.0
Age group	20-29	97(1.38)	1872(26.67)	1977(28.17)		
	30-39	17(0.24)	813(11.58)	1255(17.88)		
	40-49	19(0.27)	203(2.89)	369(5.26)		
	50-59	0	35(0.5)	91(1.3)		
	≥ 60	0	0	19(0.27)		
Educational	Primary school	3(0.04)	104(1.48)	181(2.58)	0.0129	0.0
level	Secondary school	131(1.87)	2949(42.01)	3651(52.02)		
Place of	Rural	62(0.88)	1467(20.90)	1806(25.73)	0.7176	0.06
residence	Urban	72(1.03)	1586(22.6)	2026(28.86)		
ART initiation	Pre ART	110(1.57)	2566(36.56)	2783(39.65)	<0.0001	0.0
group	Post ART	20 (24)	487(6.94)	1049(14.95)		

- The response variable (CD cell count) has 110 (1.5%) missing observations.

Moreover, from Table 3.2, there is a high prevalence of CD4 count above 500 $cells/mm^3$ for patients with low viral load at baseline (21.83%). This shows ART suppresses the amount of HIV viably in patient's body fluids who have an undetectable and low viral load at baseline to the point where standard tests are incapable of detecting any HIV or can only find a little flow. There is also a high prevalence of CD4 count above 500 $cells/mm^3$ for patients who have a stable sex partner (43.85%, p-value <0.0001) compared to patients who have many sex partners. A high prevalence of CD4 count above 500 $cells/mm^3$ is observed among patients of the age group between 20-29 years and 30-39 years, 28.17 and 17.88%, respectively (p-value< 0.0001). The prevalence of CD4 count above 500 $cells/mm^3$ is also observed among women patients with higher/secondary school levels of education (52.02%, p-value=0.0129). However, the place of residence is found not to be associated with patients' CD4 count (p-value=0.7176).

The individual profiles plot for 17 randomly selected HIV-Infected women enrolled in the CAPRISA 002 Acute Infection Study is shown in Figure 3.1.

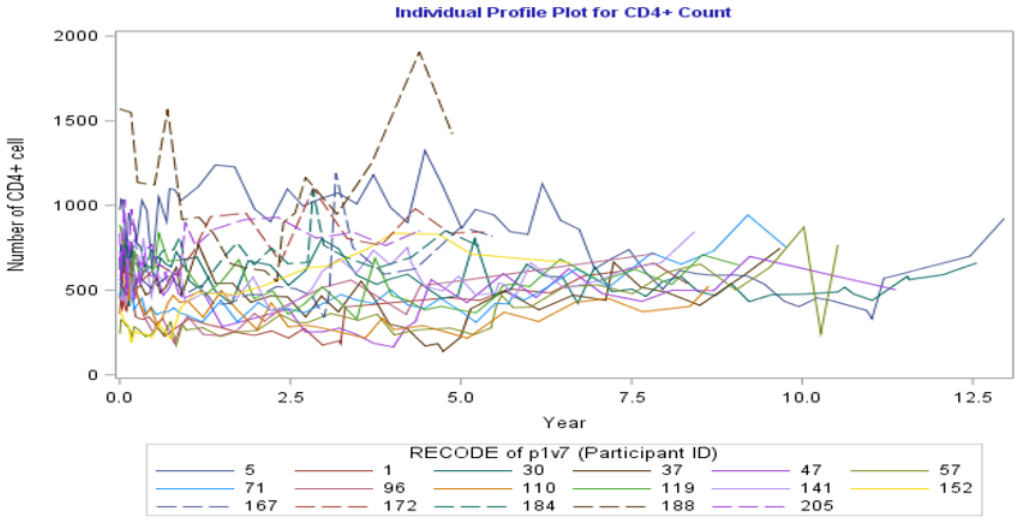


Figure 3.1: Individual Profiles plot of CD4 cell count for 17 randomly selected individuals

Analyzing data shown in Figure 3.1, we can observe insights concerning the variability between individual units at a given point in time, the variance within units over time, and the trends over time. Note that the space between the lines represents unit variability between subjects, and the change in each line (slope) represents within-subject variability. Moreover, as portrayed in Figure 3.1, CD4 cell counts appear a slightly increasing pattern over time, but the rate of increment is low. Figure 3.1 also shows that there is wide variability in the number of CD4 cells and in the number of repeated measures (number of observations per subject are not equal).

Table 3.3: Comparisons of Fit Statistics for the two distributions

	Fit statistics					
Distribution	-2 log likelihood	AIC	AICC	BIC	CAIC	HQIC
Poisson	204842.9	204892.9	204893.1	204979.4	205004.4	204927.8
Negative Binomial	**87781.28**	**87833.28**	**87833.48**	**87923.23**	**87949.23**	**87869.54**

The results of the Fit statistics in Table 3.3 are obtainable because of method=Laplace in Proc Glimmix Procedure. These values are relative and valuable when we compare different model choices. The values of the fit statistics of the NB model are much smaller than the values of the fit statistics of the Poisson model (Table 3.3). For instance, the corrected Akaike's Information Criterion (AICC) value is 87833.48 for NB versus 204893.1 for the Poisson. Also, the Pearson χ^2/DF of 20.66 for the Poisson model is problematic (Table 3.4). Ideally, this value ought to be generally 1.0 when modeling count data with a Poisson distribution. The ratio of Pearson χ^2 statistics are dropped from 20.66 to 0.91 under the NB model, which is close to one (Table 3.4), indicating that overdispersion has been appropriately modeled and it is no longer an issue under the NB model.

Table 3.4: Measure of overdispersion between Poisson and Negative Binomial distribution

Fit Statistics for Conditional Distribution	Poisson	Negative Binomial
-2 log L(CD4 counts/r. effects)	199670.3	85320.39
Pearson χ^2	145017.0	6396.89
Pearson χ^2/DF	20.66	**0.91**

In addition to the conditional fit statistics, another diagnostic that would permit us to visualize overdispersion in the Poisson model is the graphical representation (Figure 3.2). We can get residual plots through Proc Glimmix using the Plot option. Here, we only focus on looking at residual versus predicted plots. Figure 3.2 (left panel) shows the visual prove of overdispersion. As the Predicted Mean ($\hat{\mu}$) increases, the associated residuals become more broadly dispersed. The variance ought to increase as a function of the mean, but not as quickly as we see in this plot (Figure 3.2). Also, Figure 3.2 (right panel) shows prove of overdispersion. The variance adjusted residuals are more variable around the lower point of the estimated Linear Predictor ($\hat{\eta}$). On the model scale, we should not see the variance adjusted residuals variable across different points of $\hat{\eta}$ as we see in this plot (Fox & Monette, 2002; Stroup, 2012). In other words, Figure 3.2 (right panel) demonstrates that the empirical distribution of the residuals is not reasonably symmetric, and in general, it is not very informative.

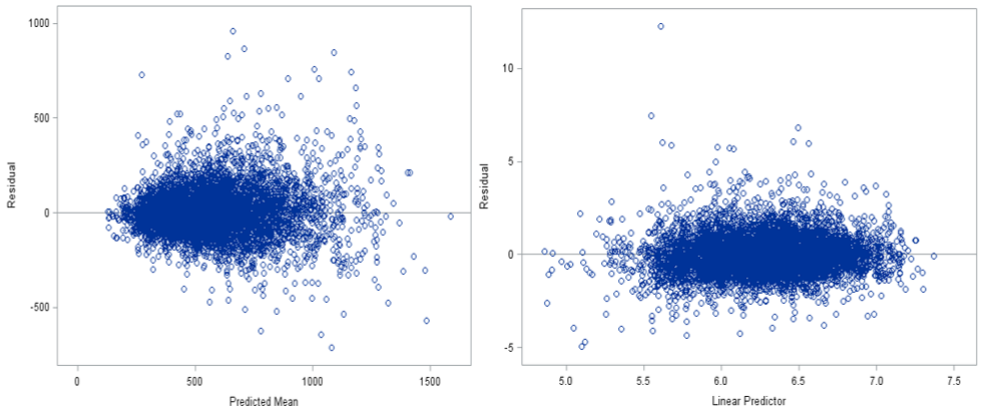


Figure 3.2: Diagnostics plot to visualize overdispersion in the Poisson regression model

The improvements in the Pearson χ^2/DF and Fit statistics indicates that it is best to model data from this experiment with the Negative Binomial distribution. Utilizing the proper distribution gives unbiased test statistics and standard error estimates.

In addition, the subsequent random effect models were taken into consideration for testing the NBMMs:
Model 1: Intercept, Time, $\sqrt{\text{Time}}$

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 t_{ij} + \beta_2 \sqrt{t_{ij}} + b_{i0} + b_{i1} t_{ij} + b_{i2} \sqrt{t_{ij}} + \varepsilon_{ij}$$

where x_{ij} is the ART initiation group variable, and t_{ij} is the time variable.

Model 2: Intercept, Time

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 t_{ij} + \beta_2 \sqrt{t_{ij}} + b_{i0} + b_{i1} t_{ij} + \varepsilon_{ij}$$

Model 3: Intercept, $\sqrt{\text{Time}}$

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 t_{ij} + \beta_2 \sqrt{t_{ij}} + b_{i0} + b_{i1} \sqrt{t_{ij}} + \varepsilon_{ij}$$

Model 4: Time, $\sqrt{\text{Time}}$

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 t_{ij} + \beta_2 \sqrt{t_{ij}} + b_{i1} t_{ij} + b_{i2} \sqrt{t_{ij}} + \varepsilon_{ij}$$

Model 5: Intercept only

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 t_{ij} + \beta_2 \sqrt{t_{ij}} + b_{i0} + \varepsilon_{ij}$$

Model 6: Time only

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 t_{ij} + \beta_2 \sqrt{t_{ij}} + b_{i1} t_{ij} + \varepsilon_{ij}$$

Model 7: $\sqrt{\text{Time}}$ only

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 t_{ij} + \beta_2 \sqrt{t_{ij}} + b_{i1} \sqrt{t_{ij}} + \varepsilon_{ij}$$

Table 3.5: Comparison of random effect models

		Information Criteria				
Random effect models	$-2\log\ell$	**AIC**	**AICC**	**BIC**	**CAIC**	**HQIC**
Model 1	**87781.28**	**87833.28**	**87833.48**	**87923.23**	**87949.23**	**87869.54**
Model 2	88603.50	88649.50	88649.66	88729.07	88752.07	88681.58
Model 3	88591.64	88637.64	88637.80	88717.21	88740.21	88669.72
Model 4	89156.39	89202.39	89202.55	89281.96	89304.96	89234.47
Model 5	89837.18	89879.18	89879.31	89951.83	89972.83	89908.47
Model 6	92302.08	92344.08	92344.21	92416.73	92437.73	92373.37
Model 7	91190.61	91232.61	91232.74	91305.26	91326.26	91261.90

We conclude that Model 1 is a preferable model among the models listed above since it has the smallest information criteria (Table 3.5). Moreover, a comparison of the covariance structure using the fitted model (Table 7.1) and a comparison of fixed-effects results across different covariance structures using Model 1 (Table 7.3) are made.

The estimated unstructured covariance matrix ($\hat{G}$) for GLMMs model that uses Negative Binomial distribution is

$$\hat{G} = \begin{bmatrix} 0.1131 & 0.000739 & -0.01754 \\ 0.000739 & 0.000155 & -0.00137 \\ -0.01754 & -0.00137 & 0.01556 \end{bmatrix}$$

The estimated scale parameter is 0.04205, which can be found in the "Covariance Parameter Estimates"output of the SAS PROC GLIMMIX (Laplace) procedure (see Table 7.2 in Appendix B). Therefore, the estimated conditional variance of the count is $\hat{\mu}_i + 0.04205\hat{\mu}_i^2$, where $\hat{\mu}_i$ is the conditional mean on the counting scale. The Scale

parameter measures the magnitude of overdispersion and is practically equivalent to the mean square error in conventional theory analysis of variance (Gbur et al., 2012).

Table 3.6 shows the overall effect of the selected factors within the fitted models. The results indicate that the effects of Time, Baseline BMI, HAART initiation group, baseline viral load, and the number of sex partners on the patient's CD4 count were highly significant in both fitted models. However, the overall F-values of the NB model were smaller than for the Poisson model. This can be supporting prove that over-dispersion can lead to inflated and biased F-values if we do not use the proper model in our analysis.

Table 3.6: Measure of over-dispersion between Poisson and Negative Binomial distribution

			NB		Poisson	
Effect	**Num DF**	**Den DF**	**F Value**	**Pr>F**	**F Value**	**Pr>F**
Time in month	1	235	62.53	<.0001	14.80	0.0002
Sqrt_Time	1	234	86.36	<.0001	48.41	<.0001
Baseline BMI category	3	6307	6.26	0.0003	6.31	0.0003
ART initiation	1	6307	345.45	<.0001	5890.28	<.0001
Baseline VL	3	6307	7.48	<.0001	12.79	<.0001
Number of sex partner	2	6307	1.64	0.1935	1.85	0.1578
Age group	5	6307	1.46	0.1987	27.34	<.0001
Education level	1	6307	0.25	0.6196	0.15	0.6990
Place of residence	1	6307	0.01	0.9246	0.11	0.7406

Table 3.7 shows the log of the expected CD4 count as a function of the selected predictor variables using a negative binomial mixed-effect model. The results indicate that time (month) significantly affects the CD4 count of a patient. We interpret the coefficient of the month as an average within-subject change in the logs of expected CD4 count for patients would be expected to increase by 0.0078 unit (p-value<0.0001; 95% CI: 0.005875, 0.009774), while holding the other factors in the model constant. The square root of time shows a significant adverse effect in the logs of expected CD4 counts of a patient (Table 3.7). Compared to pre HAART initiation, the difference in the logs of CD4 counts of a patient who had been initiated on HAART would be expected to increase by 0.2301 units (p-value<0.0001; 95% CI:

0.2058, 0.2543), holding other factors constant in the model. It can be observed that the difference in the logs of expected CD4 counts is expected to be 0.4815 units (p-value<0.0001; 95% CI: 0.2633, 0.6996) higher for patients with higher BMI (Obese) at baseline compared to patients with normal weight status holding other factors constant in the model. For those patients who had high and medium viral load at baseline, the difference in the logs of their expected CD4 counts was decreased by 0.2393 (p-value<0.0001; 95% CI: -0.3404, -0.1382) and 0.1258 (p-value=0.0061; 95% CI: -0.2157, -0.03585), respectively, compared to patients who had low viral load at baseline while holding the other factors in the model constant.

Table 3.7: Parameter estimates using Poisson and Negative Binomial mixed-effects model

			NB		Poisson	
Covariates	**Estimate**	**Std Err**	**Pr** $> \|t\|$	**95% C.I for NB Estimate**	**Estimate**	**Std Err**
Intercept	6.47	0.04982	<.0001	(6.3715, 6.5679)	6.4625	0.04264
Time in month	0.007824	0.000989	<.0001	(0.005875, 0.009774)	0.006564	0.001706
Sqrt_Time	-0.08649	0.009307	<.0001	(-0.1048, -0.06815)	-0.06839	0.009830
ART Initiation (Post)	0.2301	0.01238	<.0001	(0.2058, 0.2543)	0.1947	0.002537
Baseline BMI category (ref.=Normal weight)						
Obese	0.4815	0.1113	<.0001	(0.2633, 0.6996)	0.4985	0.1147
Overweight	0.02561	0.04975	0.6067	(-0.07191, 0.1231)	0.03131	0.05148
Underweight	0.005901	0.07927	0.9407	(-0.1495, 0.1613)	0.01691	0.08264
Baseline HIV VL category (ref.= Low VL)						
High VL	-0.2393	0.05157	<.0001	(-0.3404, -0.1382)	-0.3074	0.05065
Medium VL	-0.1258	0.04587	0.0061	(-0.2157, -0.03585)	-0.1121	0.04686
Undetectable	0.1377	0.2901	0.6351	(-0.4310, 0.7064)	0.1199	0.2978
Number of sex partner (ref.= Stable partner)						
Many partners	-0.1560	0.09394	0.0967	(-0.3402, 0.02811)	-0.1674	0.09908
No partner	-0.04821	0.04993	0.3343	(-0.1461, 0.04967)	-0.05913	0.05164
Age group in years(ref.= < 20)						
20-29	0.01166	0.03104	0.7072	(-0.04919, 0.07251)	-0.00791	0.007830
30-39	0.02852	0.03432	0.4060	(-0.03876, 0.09580)	-0.01239	0.008474
40-49	-0.00719	0.04545	0.8743	(-0.09629, 0.08191)	-0.03422	0.01112
50-59	-0.05694	0.06662	0.3927	(-0.1875, 0.07365)	-0.1399	0.01549
≥ 60	0.2082	0.1532	0.1741	(-0.09205, 0.5084)	-0.3107	0.03519
Education level (ref.= Secondary or high school)						
Primary school	-0.04509	0.09084	0.6196	(-0.2232, 0.1330)	-0.03582	0.09263
Residence of participant (ref.= Urban)						
Rural	-0.00373	0.03947	0.9246	(-0.08112, 0.07365)	0.01337	0.04038

Furthermore, the standard errors for the Poisson mixed-effects model were more likely to be underestimated and/or biased as compared to those from a negative binomial mixed-effects model since the model is fitted by ignoring overdispersion of the data (Table 3.7).

The prediction profile equation for the average number of CD4 cells following Table 3.7 results obtained by negative binomial (NB) mixed-effects model is given as:

$$\begin{aligned} log(\hat{\mu}) = 6.4697 + 0.007824 \times \text{time} - 0.08649 \times \sqrt{\text{time}} \\ +0.2301 \times \text{post HAART treatment} + 0.4815 \times \text{obese} \\ -0.2393 \times \text{high VL} - 0.1258 \times \text{medium VL} \end{aligned}$$

The prediction of individual profiles, Figure 3.3, presents the estimated trajectories for the average number of CD4 cells under the estimates obtained by the negative binomial mixed-effect model with UN covariance structure consolidated with the model where the intercept and slope were considered as random effects (see Table 7.1 and 7.3 in Appendix B) for seven patients with particular profiles for four years. For instance, from CAPRISA 002 AI Study, patient ID = 141, 22 years old female, with around 500 $cells/mm^3$ CD4 cell count at baseline, low VL at baseline, had normal weight status at baseline, and have no partner at the time of enrollment. The

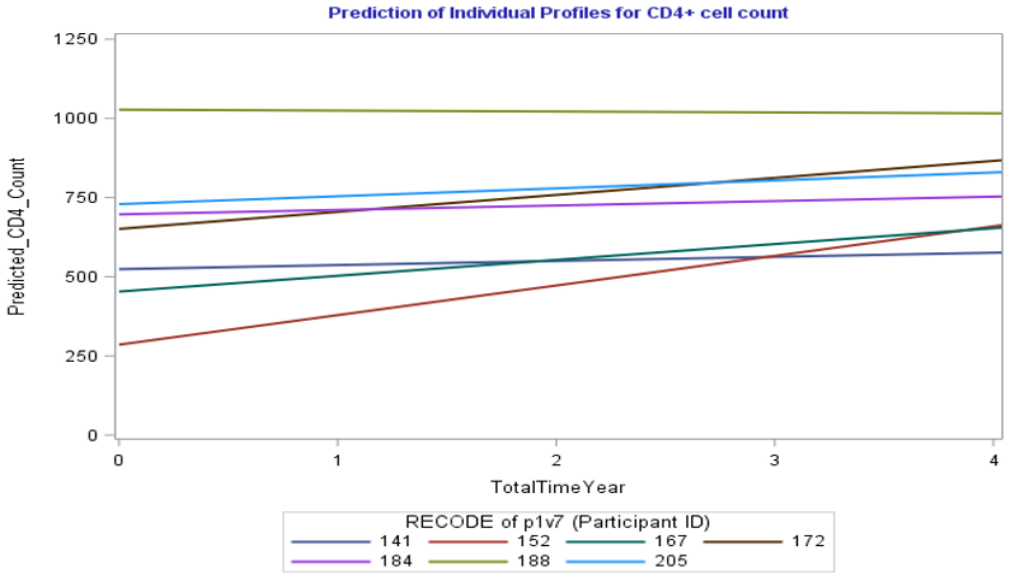


Figure 3.3: Prediction of 7 randomly selected individual profiles plot of CD4 count for four years

second patient ID=152, 34 years old female, with obese weight status at baseline, having stable sex partner, high VL at baseline, and CD4 count at baseline below 500 $cells/mm^3$. As a third example, we looked at patient ID=172 who had undetected

VL at baseline, with CD4 count at baseline above 500 $cells/mm^3$, 29 years old female, with obese weight status at baseline and have a stable sex partner. As a fourth example, we can also look at patient ID=188, who had a high number of CD4 cells at baseline (1070 $cells/mm^3$) with low VL at baseline, 42 years old, obese weight status at baseline, and have a stable sex partner. As we would anticipate, all seven individuals appeared to have an increased average number of CD4 cells over time, according to their predicted individual profiles (Figure 3.3). However, the increasing level or degree is different among individuals. This is due to factors related to this study and numerous other characteristics of these individuals, mainly (according to our research) for their VL at baseline, baseline BMI and the treatment (either the patient had effective HAART initiation after HIV exposure or not).

Moreover, for this study to yield meaningful results, we have checked the missing values in the dataset using Little's MCAR test. The regular Little's MCAR test gives us a χ^2 distance of 4515.686 with a degree of freedom 106 and p-value 0.000 (Little's MCAR test: χ^2=4515.686, DF=106, sig.= 0.000). The analysis provides evidence that the missing data in the study variables of interest are not MCAR under significance level 0.000. Therefore, we used multiple imputation (MI) techniques to get valid inferences for parameter estimates from the complete data set by fitting the chosen model. The key idea of the MI procedure is to replace each missing value with a set of m plausible values. Generally, the imputation of dependent and independent variables is basic for getting unbiased estimates of the regression coefficients (Allison, 2001). Following Rubin's (1987) terminology, the MI procedure involves three distinct phases: each missing value is imputed m times to generate m complete data sets, analyze each m complete data sets separately by using standard procedure and then combine the results to generate valid statistical inference about the model parameters from the m data set analysis using Rubin's combine rule Rubin (2004). SAS Proc MI can be used to create N number of imputations; after that, Proc MIAnalyze is used to pool the parameter estimates. A detailed discussion of missing data analysis and how missing data handled by statistical software can be found in numerous literature (Rubin, 2004; Little & Rubin, 2019; Fitzmaurice et al., 2008; Berglund & Heeringa, 2014; Enders, 2010; Molenberghs & Kenward, 2007; Bücker & Hogan, 2011; Der & Everitt, 2012).

Table 3.8 shows a combined result for each parameter. The table also displays a 95% confidence interval, the minimum and maximum regression coefficients from the imputed data set, and the associated p-value. We can compare the results given in Table 3.8 with the results of applying the negative binomial mixed-effect model to the CAPRISA 002 Acute Infection data using incomplete cases (Table 3.7). Com-

paring the two different sets of results, we do not see that many exciting differences. In both case analyses, covariates that were significantly affecting the patient's CD4 count are similar, and their respective parameter estimates are closer to each other.

Table 3.8: Combined results of a negative binomial mixed-effects model analysis using MI Procedure to deal with the missing values

	Parameter Estimates (10 Imputations)					
Parameter	**Estimate**	**Std Err**	**Pr > |t|**	**95% Confidence Limits**	**Minimum**	**Maximum**
Intercept	6.459413	0.049830	<.0001	(6.36175, 6.55708)	6.458658	6.460775
Time in month	0.007475	0.000975	<.0001	(0.00556, 0.00939)	0.007450	0.007508
Sqrt_Time	-0.083647	0.009266	<.0001	(-0.10181, -0.06549)	-0.083982	-0.083434
ART Initiation (Post)	0.224037	0.012594	<.0001	(0.19935, 0.24872)	0.223216	0.225014
Baseline BMI category (ref.=Normal weight)						
Obese	0.474714	0.109902	<.0001	(0.25931, 0.69012)	0.473892	0.475630
Overweight	0.024208	0.048971	0.6211	(-0.07177, 0.12019)	0.023820	0.024529
Underweight	0.002070	0.078101	0.9789	(-0.15101, 0.15515)	0.001321	0.003137
Baseline HIV VL category (ref.= Low VL)						
High VL	-0.239102	0.051294	<.0001	(-0.33964, -0.13857)	-0.239735	-0.238839
Medium VL	-0.122078	0.045390	0.0072	(-0.21104, -0.03311)	-0.122251	-0.121642
Undetectable	0.142848	0.286259	0.6178	(-0.41821, 0.70391)	0.142510	0.143351
Number of sex partner (ref.= Stable partner)						
Many partners	-0.153632	0.092090	0.0953	(-0.33412, 0.02686)	-0.154667	-0.152911
No partner	-0.046962	0.049227	0.3401	(-0.14344, 0.04952)	-0.047267	-0.046691
Age group in years(ref.= < 20)						
20-29	0.013477	0.031659	0.6703	(-0.04857, 0.07553)	0.012306	0.014325
30-39	0.033725	0.034974	0.3349	(-0.03482, 0.10227)	0.032678	0.034744
40-49	-0.005842	0.046177	0.8993	(-0.09635, 0.08466)	-0.007790	-0.004745
50-59	-0.052070	0.067501	0.4405	(-0.18437, 0.08023)	-0.054207	-0.051024
≥ 60	0.206708	0.156046	0.1853	(-0.09914, 0.51255)	0.205360	0.207553
Education level (ref.= Secondary or high school)						
Primary school	-0.046292	0.089605	0.6054	(-0.22191, 0.12933)	-0.046602	-0.046009
Residence of participant (ref.= Urban)						
Rural	-0.001916	0.038813	0.9606	(-0.07799, 0.07416)	-0.002146	-0.001596

In general terms, a comparison of the results from data with missing value case analysis (Table 3.7) and multiple imputation analysis (Table 3.8) shows little difference between parameter estimates, standard errors, and confidence intervals. In this case, the small difference in results and associated inferences is likely due to relatively low amounts of missing data in the analysis variables (Table 3.2). However, it

will not always be true that results from incomplete or complete case analysis and a multiple imputation treatment of the data will lead to similar results and inferences (Berglund & Heeringa, 2014). Finally, missing data is especially common in longitudinal data sets. Missingness can arise due to respondent attrition, survey structure, file-matching issues, and refusal to answer sensitive questions such as certain health conditions, illegal behaviors, or income (Berglund & Heeringa, 2014). Missing data can also arise due to death. A loss to follow-up due to death is qualitatively distinct from dropout due to other responses and, ordinarily, needs to be handled quite differently in the analysis of longitudinal data (Dufouil et al., 2004). Missing data is generally categorized as Missing Completely at Random (MCAR), Missing at Random (MAR), or Not Missing at Random (NMAR) (Rubin, 2004; Schafer, 1997; Molenberghs & Kenward, 2007; Bücker & Hogan, 2011; Raghunathan et al., 2001; Enders, 2010).

Finally, the normal probability plot of the random effects for the fitted NBMM is indicated in Figure 3.4. The assumption of normality seems reasonable for all three random effects. The plots confirmed that the estimated random effects are normally distributed, with mean zero and covariance matrix $\hat{G}$ and are independent for different group (Assumption 2 holds) (Pinheiro & Bates, 2006).

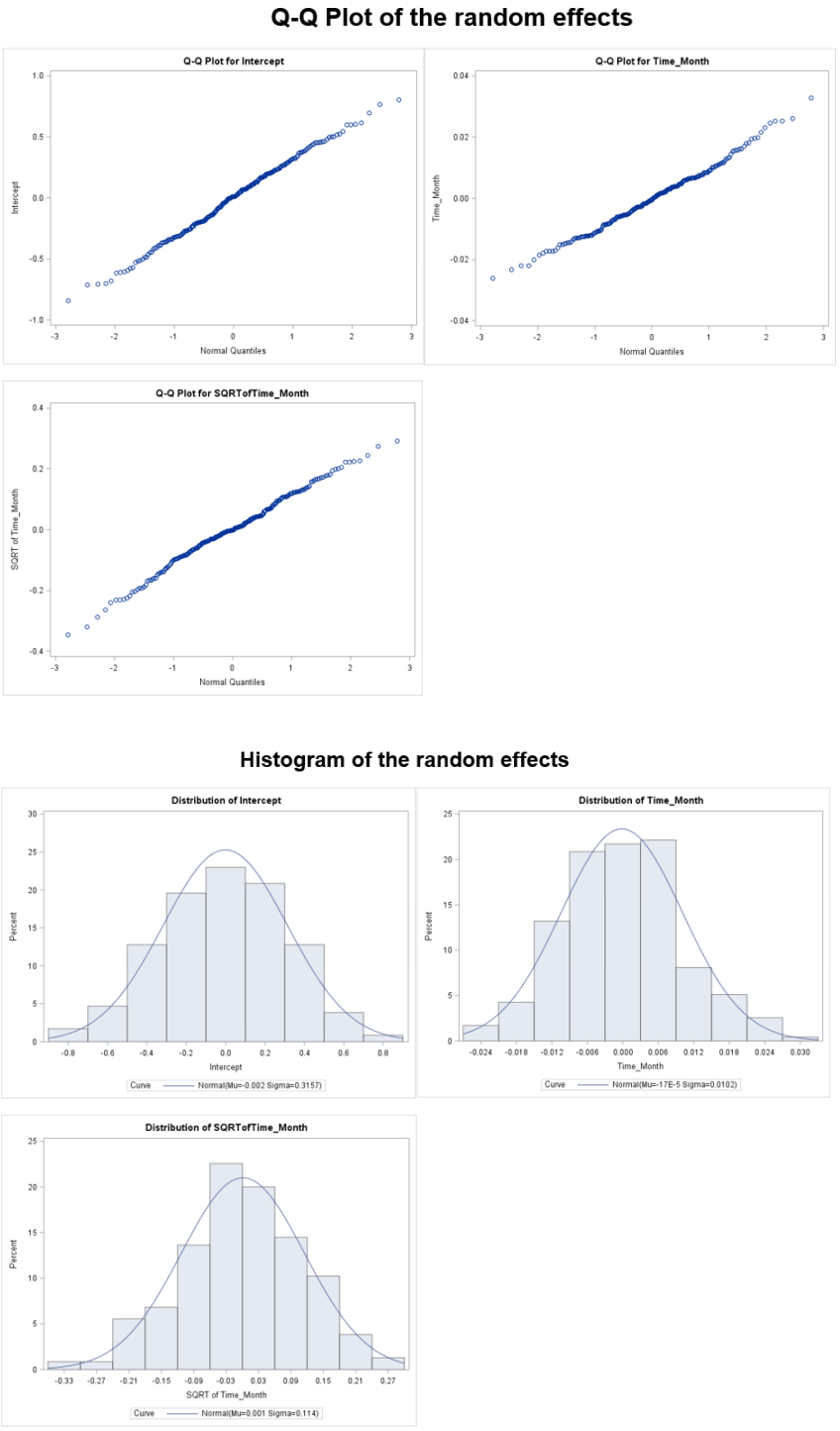


Figure 3.4: Q-Q and Histogram normal plot of the estimated random effects

Some of the codes that are used for the above section can be found here (Code 7.2 in the Appendix A).

3.8 Summary

Generalized linear models extend standard theory linear models to response variables whose distribution belongs to the exponential family. GLM comprises of three components: a *stochastic* component that characterizes the likelihood distribution of the response variable; a *linear predictor* that is a *systematic* component portraying the linear model characterized by the explanatory variables; and a *link function* that connect the mean of the response variable to a linear combination of the explanatory variables. Link functions that are commonly used for distributions are discussed in numerous literature (Dobson & Barnett, 2018; Fox & Monette, 2002; Gill & Torres, 2019; McCullagh & Nelder, 1989; Menard, 2002; Faraway, 2016; Zeger & Liang, 1986; Rawlings et al., 2001; Stroup, 2012; Jiang, 2007). Parameters in GLM are estimated based on maximum likelihood principles. Different ways of transformations of the response variable make the transformed data fulfill the linear model's assumptions, such as approximately normally distributed and having stable variances. In a more common term, a transformation is a replacement that changes the shape of distribution or relationship. However, transformation can be problematic for regression settings in which it also influences the functional relationship between the explanatory and the outcome variable. Sometimes it is not recognized that the use of transformations changes the model under consideration (McArdle & Anderson, 2004).

Transformations can be problematic when a specific choice is not predetermined by other considerations; that is, the choice of transformation is subjective (Mahmud et al., 2006). GLMs avoid these issues since the data are not transformed; instead, a function of the means is modeled as a linear combination of the explanatory variables (Gill & Torres, 2019; McCullagh & Nelder, 1989). In some cases, for example, for large values of the estimated coefficient, the use of a transformation is effective than using GLMs and Wald type statistics for inference (Menard, 2002; Rawlings et al., 2001). In general, however, transformations rarely compete well with GLMs for adequately powered studies (McCullagh & Nelder, 1989). Therefore, we analyzed the non-normal untransformed form of the CD4 cell count of a patient enrolled in the CAPRISA 002 Acute Infection Study in the context of generalized linear mixed-effects models (Table 3.7).

Longitudinal studies, also called mixed-effects models, are used to study changes in the response variable over a relevant interval of time or space and the effects of different factors on these changes (Yirga et al., 2020b). The two fundamental issues in longitudinal studies are constructing an appropriate model for the mean and selecting an adequate but parsimonious model for the covariance structure of longi-

tudinal data (Fitzmaurice et al., 2012; Yirga et al., 2020b). For these reasons, we have fitted a negative binomial mixed-effects model consolidated with the UN covariance structure since there was enough evidence of overdispersion in the data, and the chosen covariance structure gives the smallest information criteria (Table 7.1 in the Appendix B). The comparisons between Poisson and negative binomial mixed-effects models were outlined in Table 3.7. Moreover, comparisons of the covariance structure are illustrated in Table 7.1 in the Appendix B). Generalized linear mixed-effects models combine the generalized linear models with the linear mixed models. As an extension of generalized linear models, they consolidate random effects into the linear predictor. As a mixed model, they contain at least one fixed effect and at least one random effect. Parameter estimation in GLMMs is also based on maximum likelihood principles; inferences for the parameters are readily obtained from classical maximum likelihood theory (McCulloch & Neuhaus, 2014; Fitzmaurice et al., 2012; Yirga et al., 2020b). The two fundamental computational approaches to obtain solutions to the likelihood equations are a pseudo-likelihood and integral approximation of the log-likelihood using either the Laplace or Gauss-Hermite quadrature strategies (Der & Everitt, 2012; Shoukri, 2018; Stroup, 2012). Since pseudo-likelihood produces biased covariance parameter estimates when the number of observations per subject is small, it is especially inclined to biased estimates when the power is small and uses a pseudo-likelihood rather than a true likelihood, likelihood ratio and fit statistics such as AICC and BIC have no clear meaning (Yirga et al., 2020b). However, the Laplace and quadrature approaches use the actual likelihood and grant us the appropriate likelihood ratio tests or information criteria, permitting competing models to be compared using these test statistics. Of these two, the Laplace method is best since quadrature is ordinarily computationally restrictive for regularly repeated measures. Moreover, the Laplace procedure is less computationally intensive than the quadrature procedure and is considerably more flexible in terms of the models with which it can be used (Yirga et al., 2020b). Detailed discussions of parameter estimation in GLMMs can be found in numerous literature (Stroup, 2012; Faraway, 2016; Fitzmaurice et al., 2012; McCullagh & Nelder, 1989; Zeger & Liang, 1986; Jiang, 2007). The fit statistics in Table 3.4 were obtained by using the Laplace method. If this method had not been specified on the SAS Proc Glimmix procedure, the default pseudo-likelihood procedure would have been used to fit the model. Because pseudo-likelihood is based on Tylor series approximation to the conditional likelihood and not explicitly on the conditional likelihood itself, a goodness of fit statistic such as the Pearson χ^2 that is particularly appropriate to the conditional distribution cannot be computed. Instead, the pseudo-likelihood approaches calculate a Generalized χ^2 statistic that measures the combined fit of the conditional distribution of the counts and the random effects. Since it is not particular to only

the conditional distribution, it does not give a clear-cut diagnostic to evaluate the fit of the Poisson distribution to the counts (Der & Everitt, 2012).

The Pearson χ^2/DF gives the goodness of fit statistic to evaluate over-dispersion within the Poisson model. Since the mean and variance of the Poisson are equal, the scale parameter (α) is one. If the Poisson assumption is fulfilled, the Pearson χ^2/DF ought to be close to one. Its estimated value of 20.66 (Table 3.4) indicated solid prove of overdispersion under the Poisson model (Yirga et al., 2020b). Overdispersion would imply more variability shown by the data than would be assumed under a given statistical model (Morel & Neerchal, 2012; Yirga et al., 2020b). Overdispersion could be an issue that should not be disregarded in the analysis. The essential and most serious consequence of overdispersion is its effect on standard errors and test statistics. This was illustrated in Table 3.6, uncorrected analysis of overdispersed data (Poisson model) results in underestimated standard errors, leading to biased estimates and inflated test statistics. It is basic to check for overdispersion when fitting a GLM or a GLMM to guarantee that inferences derived from the fitted model are precise (Morel & Neerchal, 2012; Yirga et al., 2020b). Overdispersion is an implication that the fitted model is incorrect, and adjustments are required. The two most commonly used approaches in GLMMs, to avoid unwanted outcomes outlined above, are: adjusting the standard errors and test statistics by incorporating an adjustment for overdispersion in the model or assume a different probability distribution for the counts that more reasonably approximate the method by which overdispersion emerge (McCullagh & Nelder, 1989). Because the second strategy of assuming a different distribution is a reasonable and suggested methodology, it was illustrated in Table 3.5 in which the negative binomial distribution substitutes the Poisson distribution as the conditional distribution of the outcome. The NB distribution is the foremost candidate as an alternative to the Poisson (Hilbe, 2011; Yirga et al., 2020b). The Pearson χ^2/DF value of 0.91 (Table 3.4) shows that the negative binomial provides a much-improved fit of the data compared to the Poisson model. This is one of the reasonable GLMMs approaches for managing with overdispersion.

Table 7.3 outlined that the fixed effects results can be significantly influenced by the covariance structure. Furthermore, the covariance structure also impacted the estimate of the random effects: the time effects and their standard errors. The standard errors tend to be affected more than the estimates. The choice of covariance structures matters for non-normally distributed data, just as it does for normally distributed data. The fit statistics associated with pseudo-likelihood estimation are not comparable among models. Consequently, the fit statistics cannot be used to select between competing for covariance structures. Therefore, the choice of covariance

structure is not as straightforward for non-normal longitudinal response data as it is under normality assumption (Harrison et al., 2018; Shoukri, 2018; Stroup, 2015; Gbur et al., 2012; McArdle & Anderson, 2004; Galecki, 1994). However, for the GLMM approach, the situation is better. As we discussed previously since the GLMM characterizes an exact probability process under the Laplace method, fit statistics such as AICC and BIC can be obtained (Harrison et al., 2018). Thus, for GLMMs, covariance structures selection can proceed much as it does for normally distributed data as long as either Laplace (preferable) or quadrature methods are used. Moreover, while we have incorporated a parametric spatial covariance structure for the fitted negative binomial mixed-effects model, other approaches to account for spatial variation are of interest. Our study methodology, in theory, can be extended to address this issue using a generalized linear mixed-effects model for spatial data (Schabenberger & Gotway, 2017). Therefore, we leave this and other conceivable extensions for future research.

Along this line, it would be fascinating to extend this study to the quantile mixed-effects model. The majority of longitudinal modeling methods are based on mean regression to concentrate only on the average effect of covariate and the mean trajectory of the longitudinal outcome, which is constant across the population (Yirga et al., 2020b). However, such average effects are not always of interest in many areas and sometimes quite heterogeneous. Thus, the quantile mixed-effects model has the capacity, at both the population and individual level, to identify heterogeneous covariates effects and describe differences in longitudinal changes at different quantiles of the outcome. Hence leads to more efficient estimates, especially when the errors are over-dispersed (Koenker, 2004; Geraci & Bottai, 2014).

Chapter 4

Application of quantile mixed-effects model in modeling CD4 count from HIV-infected patients in KwaZulu-Natal South Africa

4.1 Introduction

The classical regression model has been the commonly applied statistical procedure to depict the effects of explanatory variables on the mean response. This traditional regression assumes that the effects of covariates are the same throughout the population. Nonetheless, such results based on a fixed location may not be relevant in numerous areas, and sometimes the community is entirely diverse. Numerous investigators, economic experts, monetary stakeholders, clinicians, and legislators have revealed a growing interest in group differences across the whole population instead of entirely depending on the average (Davino et al., 2013; Girma & Görg, 2005; Chunying, 2011; Mirnezami et al., 2012). Conventional regression cannot satisfy all of these demands or conditions. In mean regression, we can only study the influence of independent variables on the conditional mean of the outcome. Another approach to study the central location is median regression. The median regression approach is vigorous to the manifestation of outliers, and when the error distribution is not correctly specified (Davino et al., 2013; Koenker & Bassett, 1978).

Quantile regression (QR) Was popularized by Koenker & Bassett (1978). It is an extension of median regression to examine the covariates' influence on different quantiles/percentiles or the entire response distribution. Fixed effects could have different impacts across various quantile levels. QR allows for a wide-ranging of applications; for example, investigating the 5^{th} or 25^{th} percentile (lower quantiles) of the response (e.g., CD4 count distribution of HIV infected patient) might be of interest in studying patients with fewer CD4 cell counts, where individuals are at higher risk of developing illnesses. Therefore, that will be a high qualification for immediate HAART treatment so that the patients become beneficial. Therefore, studying the response across all quantiles (e.g., at different CD4 count distribution), rather than only the central tendency, as in mean regression, is important. The central tendency cannot represent the entire distribution.

In recent years, mixed quantile regression models have become a widely used technique in statistical studies. A quantile regression model is based on conditional quantiles instead of modeling the effects of covariates on the conditional mean, which extends regression for the mean to the conditional distribution of the outcome variable. Therefore, it is possible to examine the location, scale, and shape of the distribution of responses to get an idea of how the covariates affect the distribution of responses. It is also more robust to outliers when compared to conventional mean regression and is invariant to monotonic transformations. There is no need to make any Gaussian assumptions concerning the response with a quantile regression, and it is capable of handling heavy-tailed and asymmetric data. As a result, CD4 count can be modeled very well using this method.

Data gathered in numerous longitudinal research register considerable information on repeated measures and imperative for understanding disease progression in clinical studies. For instance, in HIV/AIDS investigations, repeated counts of CD4 cells are vital signs of the seriousness of the viral infection, disease development, therapy assessment and can be used to detect the future advantages of medical involvement and risk factors for poor outcomes. Mixed-effects modeling have become quite popular in practical statistics. They are often used to examine longitudinal data due to their ability to deal with both between-subject and within-subject variability in longitudinal data (Pinheiro & Bates, 2006). Mixed-effects models and their estimated effects are formulated on the response variable via a mean regression, regulating between-subject heterogeneity through normally distributed subject-specific random effects and random errors. But, this centrality-based inferential system is regularly cannot represent the entire distribution and may not be the finest location to characterize the data. For more details on mixed-effects models (see Pinheiro &

Bates, 2006; Verbeke & Molenberghs, 2009; Twisk, 2013; Diggle et al., 2002; Brown & Prescott, 2014). There are also various strategies applicable to handle longitudinal data, for instance, mixed-effects models and generalized linear mixed-effects models, as we have discussed in the previous chapters. However, all these techniques limit the investigation of variations between subjects with regard to the mean of the response variable, and they utilize parametric models based on the distributional hypothesis (Davino et al., 2013). Moreover, in some cases, it could be challenging to obtain appropriate transformation to normality for the response variable, or some objection to outliers may be required. A good solvent to all these matters is given by concentrating on the conditional quantiles of the longitudinal outcome (Koenker, 2005b). "Conditional QR methods, measuring the complete conditional distribution of the response variable, have been developed to grant an analysis of variable effects at any subjective quantiles of the response distribution. Furthermore, QR techniques do not require any distribution assumption on the error; besides that, the error term has a zero conditional quantile, like the ALD" (Wichitaksorn et al., 2014).

4.2 Quantile Regression

Quantile regression (QR) is a cutting-edge statistical strategy for modeling the percentiles of a response variable conditional on explanatory covariates. While regression for medians can be seen as more robust than regressions to model the mean value, QR, a generalization of median regression, enables more fully to explore the data by modeling the conditional quantiles at low or high quantiles, such as the 5th and 95th percentiles. Studying the entire distribution of the response rather than only the central tendency, as in mean regression, is important. Especially when the distribution has a heteroscedastic nature, only the central tendency cannot represent the entire distribution. If most of the observations are concentrated, for instance, on the 75^{th} percentile of the distribution, then it is more appropriate to consider the 75% regression quantile than mean regression. Further, QR does not assume a specific form for the (conditional) distribution and thus is able to accommodate non-normal errors. When the response variable given a set of covariates has a heavy-tailed distribution, QR puts a reduced weight on the extreme observations. Furthermore, due to its robustness to outliers, there is a growing interest in the literature on quantile regression. For these reasons, QR becomes more prevalent in clinical, biomedical, and other health-related research. For instance, Yirga et al. (2018) examined the BMI of under-five children as a function of age and other important factors by quantile regression. More applications of QR to independent data can be found in a number of areas, among which public health, bioinformatics, health care, environmental science, ecology, microarray data analysis, and survival data analysis (Koenker, 2005b;

Buchinsky, 1998; Ellerbe et al., 2013; Koenker & Hallock, 2001; Peterson & Krishnan, 2015; Song et al., 2017; Sherwood et al., 2013; Cook & Manning, 2009; Borgoni, 2011; Yu et al., 2003; Knight & Ackerly, 2002; Cade & Noon, 2003).

QR allows us to look beyond the average and provide a description of the whole conditional distribution of a response variable in terms of a set of explanatory variables. It offers, therefore, an invaluable tool to discern effects that would be otherwise lost in the conventional regression model analyzing the sole conditional mean (Davino et al., 2013). Conventional regression focuses on the expectation of variable Y conditional on the values of a set of variables $\boldsymbol{x}$, $E(Y|\boldsymbol{x})$, the so-called regression function (Gujarati, 2014; Weisberg, 2005). Such a function restricts exclusively on a specific location of Y conditional distribution. QR extends this approach by allowing one to study the conditional distribution of Y on $\boldsymbol{x}$ at different locations and thus offering a global view on the interrelations between Y and $\boldsymbol{x}$.

QR solutions are computed for a selected number of quantiles, typically the three quantiles along with two extreme quantiles, that is for

$$\tau = \{0.05, 0.25(Q_1), 0.5(Q_2), 0.75(Q_3), 0.95\}.$$

This is in light of the search for a rightful compromise between the amount of output to manage and the results to interpret and summarize (Davino et al., 2013). Although in many practical applications of QR, the focus is on estimating a subset of quantiles, it is worth noticing that it is possible to obtain estimates across the entire interval of conditional quantiles. In particular, the set: $\{\boldsymbol{\beta}_{(\tau)} : \tau \in (0,1)\}$ is referred to as the quantile process (Koenker, 2005b).

Estimation of conditional quantiles relies on the non-differentiable and asymmetric loss(check) function of Koenker & Bassett (1978), $\rho_\tau = u(\tau - I\{U < 0\})$ for $\tau \in (0,1)$, with $I\{\cdot\} = 1$ if the function holds, and 0 otherwise, rather than the square loss function in mean regression (Koenker, 2005b). Therefore, the computational and theoretical aspects of conditional QR are different to that of mean regression.

4.2.1 Unconditional quantiles

Consider the mean of a generic r.v. Y, denoted as $\mu = E(Y)$ which is defined as the center c of a univariate distribution that minimizes the squared sum of deviations:

$$\mu = \arg\min_c E(Y - c)^2 \tag{4.1}$$

The median (M), which is the middle value (or the value half-way between the two middle values) of a set of ranked data, rather, minimizes the absolute sum of deviations:

$$M = \arg\min_{c} E|Y - c| \tag{4.2}$$

A particular location of the distribution (univariate quantile), in other words the τ^{th} quantile, is the value of y such that $P(Y \leq y) = \tau$. As a starting point, consider the cumulative distribution function (cdf): $F_Y(y) = F(y) = P(Y \leq y)$. Thus, the τ^{th} quantile of Y, denotes as $Q_\tau(Y)$, is defined as the inverse of the cdf: $Q_\tau(Y) = Q_\tau = F_Y^{-1}(\tau) = \inf\{y : F(y) \geq \tau\}$, for $\tau \in (0, 1)$. If $F(\cdot)$ is strictly increasing and continuous, then F_Y^{-1} is the unique real number y such that $F(y) = \tau$ (Gilchrist, 2000). Therefore, $Q_\tau(Y)$ minimizes the expected check function:

$$= \arg\min_{c} E[\rho_\tau(Y - c)], \tag{4.3}$$

In such a view the median regression, which is a special case of quantile regression with $\tau = 0.5$, can bewritten as:

$$Q_{0.5}(Y) = \arg\min_{c} E[\rho_{0.5}(Y - c)] = \arg\min_{c} E[0.5|Y - c|],$$

where $\rho_\tau(\cdot) = [\tau - I(y < 0)]y = [[(1 - \tau)I(y \leq 0) + \tau I(y > 0)]|y]$ (Davino et al., 2013). This check (loss) function is then an asymmetric absolute loss function that is a weighted sum of absolute deviations, where a $(1-\tau)$ weight is assigned to the negative deviations and a τ weight is used for the positive deviations. The loss function ρ_τ of Koenker & Bassett (1978) is non-differentiable at zero. Thus, the minimizer has no explicit solution. This calls for the use of optimization methods such as Linear programming (LP).

Linear Programming

The problem which seeks to optimize a given linear function subject to linear equations and inequalities is called *linear program* (Koenker, 2005b; Vanderbei, 2020; Matousek & Gärtner, 2007). *Linear programming* (LP) is a subset of mathematical programming, facing the efficient allocation of limited resources to know activities with the objective of meeting the desired goal. For instance, let the random variables: $x_i \geq 0, i = 1, ..., n$, whose values are to be decided in some optimal fashion, are referred to as *decision variables* (Vanderbei, 2020; Davino et al., 2013). Hence, LP aims to find a vector $\boldsymbol{X}^* \in \Re_+^n$ to minimizing (or maximizing) the value of a given linear function among all vectors $\boldsymbol{X} \in \Re_+^n$ that satisfy a given system of linear equations

and inequalities. This linearity has two purposes: to measure the considered quantities with a linear function and to restrict the feasible plans by linear constraints (inequalities) (Davino et al., 2013).

LP is a flexible approach widely used in various studies with different aims. An exhaustive treatment of the topic would be outside the scope of this thesis. However, detailed description of the classical theory of LP as well as its formulation for simple, multiple, and quantile regression models can be found Matousek & Gärtner (2007), Koenker (2005b), Davino et al. (2013), or Vanderbei (2020).

4.2.2 Conditional quantiles

The idea of the unconditional mean as the minimizer of Equation (4.1) can be extended to the estimation of the conditional mean function by incorporating the effect of covariates, $\boldsymbol{X}$, on the response variable, Y:

$$\hat{\mu}_{(\boldsymbol{X},\boldsymbol{\beta})} = \arg\min_{\mu} E[Y - E(Y|\boldsymbol{X} = x)]^2, \tag{4.4}$$

where $E[Y - E(Y|\boldsymbol{X} = x) = \boldsymbol{X}'\boldsymbol{\beta}$ in the case of a linear mean function. Thus, the coefficient vector $\boldsymbol{\beta}$ is obtained by rearranging the above equation becomes: $\hat{\boldsymbol{\beta}} = \arg\min_{\beta} E[Y - \boldsymbol{X}'\boldsymbol{\beta}]^2$. Proceeding similarly, the τ^{th} conditional quantile of Y, denoted as $Q_\tau(Y|\boldsymbol{X})$ is obtained as:

$$\hat{Q}_\tau(Y|\boldsymbol{X}) = \arg\min_{Q_\tau(Y|\boldsymbol{X})} E[\rho_\tau(Y - Q_\tau(Y|\boldsymbol{X}))] \tag{4.5}$$

where the (τ)-notation denotes that the parameters and the corresponding estimators are for a particular quantile τ (Davino et al., 2013).

As mentioned previously, QR is an extension of the conventional estimation of conditional mean models to conditional quantile functions; that is, an approach allowing us to estimate the conditional quantiles of the distribution of a response variable Y in the function of a set of predictor variables $\boldsymbol{X}$. In the framework of linear regression, the QR model for a given conditional quantile τ can be formulated as follows:

$$Q_\tau(Y|\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}_\tau, \tag{4.6}$$

where $0 < \tau < 1$ and $Q_\tau(\cdot|\cdot)$ denotes the conditional quantile function for the τ^{th} quantile.

The quantile level is frequently signified by the Greek letter τ. Quantiles are location

and scale parameters at the same time. For a given $\tau \in (0,1)$, the τ^{th} quantile is the value of a r.v, where $\tau \times 100\%$ of its value lie below. In other words, it is the value such that at most $(1-\tau) \times 100\%$ of the values lies above. Thus, τ^{th} quantiles close to 0.5-quantile give the median, which is a well-known location parameter. On the other hand, τ^{th} quantiles close to zero or one give an idea of the scale. For instance, the interquartile range (IQR) is defined as the 0.75-quantile minus the 0.25-quantile: IQR=$Q_3 - Q_1$.

Let y denote a scalar response variable with conditional cumulative distribution function F_y, whose shape is unspecified and x_i the corresponding covariates vector of dimension $k \times 1$ for subject $i, i = 1, ..., n$. Then, following Koenker & Bassett (1978), the τ^{th} conditional quantile regression model is written as $Q_\tau(y_i|\boldsymbol{x}_i) = \boldsymbol{x}_i'\beta_\tau$, where $Q_\tau(y_i|\boldsymbol{x}_i) \equiv F_{y_i}^{-1}(\cdot)$, which is the quantile function (or the inverse cumulative distribution function) of y_i given x_i estimated at τ, and β_τ is a column vector of regression parameters corresponding to the τ^{th} quantile. On the other hand, this expression can be written as

$$Q_\tau(y_i|\boldsymbol{x}_i) = \boldsymbol{x}_i'\beta_\tau + \varepsilon_i, \qquad \text{with} \quad Q_\tau(\varepsilon_i|\boldsymbol{x}_i) = 0, \tag{4.7}$$

where ε_i is the error term whose distribution (with density $f_\tau(\cdot)$) is restricted to have the τ^{th} quantile to be zero, that is $\int_{-\infty}^{0} f_\tau(\varepsilon_i)d\varepsilon_i = \tau$ (Liu & Bottai, 2009; Lachos et al., 2015). "The error density $f_\tau(\cdot)$ is often left unspecified in the classical literature"(Lachos et al., 2015). Thus, the estimator $\hat{\boldsymbol{\beta}}_\tau$ proceeds through *linear programming* (LP) by minimizing

$$\hat{\beta}_\tau = \underset{\beta \in R^P}{\arg\min} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i'\beta_\tau) \tag{4.8}$$

where $\rho_\tau(\cdot)$ is the so called *loss (or check) function* defined by $\rho_\tau(u) = u(\tau - I\{u < 0\})$ with u being a real number and $I\{\cdot\}$ indicates the indicator function. Thus, $\hat{\beta}_\tau$ is called the τ^{th} quantile regression estimate (Koenker & Bassett, 1978; Koenker & Machado, 1999; Koenker & Hallock, 2001; Koenker, 2005b). For a special case of $\tau = 0.5$, corresponds to median regression, Equation (4.8) simplified to

$$\hat{\beta}_{0.5} = \underset{\beta \in R^P}{\arg\min} \sum_{i=1}^{n} |y_i - \boldsymbol{x}_i'\beta_{0.5}|$$

The parameter β_τ and its estimator $\hat{\beta}_\tau$ depends on the quantile τ, due to the fact that different choices of τ estimates different values of β (Liu & Bottai, 2009). For this reason, interpretation of β_τ is specific to the quantile being estimated, the inter-

cept term denotes the baseline predicted value of the response at specific quantile τ, while each coefficient can be interpreted as the rate of change of the τ^{th} response quantile per unit change in the value of the corresponding predictor variable (i^{th} regressor) keeping all the other covariates constant: $\beta_\tau = \frac{\partial Q_\tau(y_i|\boldsymbol{x}_i)}{\partial x_i}$.

If the distribution of the error term ε, characterized by its distribution function F_ε is known, Equation (4.7) would be stated as follows

$$Q_\tau(y_i|\boldsymbol{x}_i) = \boldsymbol{x}_i'\beta_\tau + F_\varepsilon^{-1}(\tau), \qquad i = 1, ..., n.$$

In this case, the conditional mean and other associated measures of dispersion could have better properties; therefore, there is no need for QR under these models. However, knowing the distribution of the error term ε in real data analysis is rare; rather, long-tailed errors or heteroscedastic models or a mixture of both are usually observed. For these reasons, either a robust alternative model (Koenker & Bassett, 1978) or a heteroscedastic extension of the model (Searle, 1997; Searle & Gruber, 2016) is needed.

The objective function of the conditional quantile estimator, $\hat{\boldsymbol{\beta}}_\tau$, in Equation (4.7) proceeds by minimizing

$$\begin{aligned} H(\beta_\tau) &= \sum_i \tau|\varepsilon_i| + \sum_i (1-\tau)|\varepsilon_i| \\ &= \sum_{i:y_i \geq \boldsymbol{x}_i'\beta_\tau}^{n} \tau|y_i - \boldsymbol{x}_i'\beta_\tau| + \sum_{i:y_i < \boldsymbol{x}_i'\beta_\tau}^{n} (1-\tau)|y_i - \boldsymbol{x}_i'\beta_\tau|, \qquad 0 < \tau < 1 \end{aligned} \tag{4.9}$$

where $i : y_i \geq \boldsymbol{x}_i'\beta_\tau$ for under prediction, $i : y_i < \boldsymbol{x}_i'\beta_\tau$ for over prediction, and $\hat{\boldsymbol{\beta}}_\tau$ is the point where the absolute distance of all observations below are weighted with $1-\tau$ and the ones above are weighted with τ (Koenker & Bassett, 1978). Since the above objective function is nondifferentiable, the gradient optimization methods are not applicable; instead, linear programming methods can be used to obtain $H(\beta_\tau)$ (Cameron & Trivedi, 2005; Cameron et al., 2009). As discussed under Cameron et al. (2009), the quantile regression estimator is asymptotically normal under general conditions, and it is given by

$$\hat{\boldsymbol{\beta}}_\tau \overset{a}{\sim} N(\boldsymbol{\beta}_\tau, \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}),$$

where $\boldsymbol{A} = \sum_i \tau(1-\tau)\boldsymbol{x}_i\boldsymbol{x}_i'$, $\boldsymbol{B} = \sum_i f_{u_\tau}(0|\boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i'$, and $f_{u_\tau}(0|\boldsymbol{x}_i)$ is the conditional density of the error term $u_\tau = y - x'\beta_\tau$ evaluated at $u_\tau = 0$ Buchinsky (1998). Estimation of the variance of $\hat{\boldsymbol{\beta}}_\tau$, which involves $f_{u_\tau}(0|\boldsymbol{x}_i)$ to estimate, is complicated

and computationally intense. However, statistical software packages such as *SAS Proc Quantreg*, *quantreg* in R, and *qreg*, *bsreg*, *sqreg* in Stata can easily obtain standard errors for $\hat{\boldsymbol{\beta}}_\tau$.

One of the vital properties of conditional quantiles is their behavior with respect to monotone transformations of the response variable. Transforming the covariates such that they have equivalent scales expecting such changes have no fundamental changes on the coefficients referred to as *equivariance* (Buchinsky, 1998).

The conditional quantiles' equivariance properties

Consider the simple QR model with one explanatory variable for a given quantile τ: $Q_\tau(\hat{y}|\boldsymbol{x}) = \hat{\beta}_{0(\tau)} + \hat{\beta}_{1(\tau)}x$. For $\tau \in (0,1)$, the equivariance property for chosen transformation can be written as:

- **Scale equivariance**: $Q_\tau(c\hat{y}|\boldsymbol{x}) = c\hat{\beta}_{0(\tau)} + c\hat{\beta}_{1(\tau)}\boldsymbol{x}$,
  $:Q_\tau(d\hat{y}|\boldsymbol{x}) = d\hat{\beta}_{0(\tau)} + d\hat{\beta}_{1(\tau)}\boldsymbol{x}$,
  where c and d denote a positive and negative multiplier constant, respectively. As a special case, $\tau = 0.5(Q_2)$ the QR estimates are scale equivariant, irrespective of the sign of the constant (Manning et al., 1998; Davino et al., 2013).
- **Shift or regression equivariance**: $Q_\tau(\hat{y}^*|\boldsymbol{x}) = \hat{\beta}_{0(\tau)} + [\hat{\beta}_{1(\tau)} + \gamma]\boldsymbol{x}$,
  where the dependent variable y is obtained as a linear combination, through the γ coefficient, of the explanatory variable. Such an effect holds when y is subjected to a location shift (Kuan, 2007): $y^* = y + x\gamma$.
- **Equivariance to reparametrization of design**: $Q_\tau(\hat{y}|\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A}^{-1}\boldsymbol{x}\hat{\beta}_{(\tau)}$,
  where $\boldsymbol{A}$ be a $p \times p$ non-singular matrix.
- **Equivariance to monotone function**: $Q_\tau(h(\hat{y})|\boldsymbol{x}) = h(\hat{\beta}_{0(\tau)}) + h(\hat{\beta}_{1(\tau)})\boldsymbol{x}$,
  where $h(\cdot)$ is a non-decreasing function on $\Re$.

Remark: the monotone transformation property is peculiar to QR, while the first three properties are also satisfied by OLS estimators (Manning et al., 1998; Koenker & Bassett, 1978; Davino et al., 2013).

Equivariance to monotone transformation is vital in real data applications because the appropriate selection of the $h(\cdot)$ monotone function is necessary to manage and correct different kinds of skewness (Manning et al., 1998; Davino et al., 2013). For instance, the logarithmic transformation is a non-decreasing function that is typically applied when the variable is right-skewed. Such a transformation can only be used in the case of positive values (Davino et al., 2013).

The case of a log transformation of Y for the above equivariance to monotone function can be expressed as:

$$Q_\tau(\log(\hat{y})|\boldsymbol{x}) = \log(\hat{\beta}_{0(\tau)}) + \log(\hat{\beta}_{1(\tau)})\boldsymbol{x}.$$

However, $\log(E(\hat{y}|\boldsymbol{x})) \neq E(\log(\hat{y})|\boldsymbol{x})$, in the case of OLS regression. The log transformation might be very hazardous in terms of the inference results of an OLS regression (Manning et al., 1998) whereas it may aid the statistical inference at QR (Cade & Noon, 2003).

In OLS regression, inference on a transformed dependent variable should be interpreted very cautiously because the evaluation of the significance of the parameter values can lead to different conclusions with and without the use of the transformation. However, inference on the QR results is not affected by a monotone transformation, and it can even be improved (Chen, 2005). More practical examples of evaluating the consequence of a log transformation for the parameter inference in detail, both in OLS and QR analysis, can be found in Davino et al. (2013). The equivariance property of QR to a different monotone transformation capable of dealing with negative skewness is presented by Manning et al. (1998). Moreover, equivariance properties and their corresponding proofs are presented by Koenker & Bassett (1978).

4.2.3 Asymmetric Laplace Distribution for Quantile Regression

As illustrated in Koenker & Bassett (1978), the check function $(\rho_\tau(\cdot))$ is not differentiable at zero; thus specific solutions to the minimization problem cannot be extracted. Hence, LP procedures are often used to achieve a relatively fast computation of $\hat{\boldsymbol{\beta}}_\tau$ (Lachos et al., 2015; Cameron & Trivedi, 2013). A natural link between minimization of the quantile check function and ML theory is given by the assumption that the error term in Equation (4.7) follows an asymmetric Laplace distribution (ALD) (Yu & Moyeed, 2001; Koenker & Machado, 1999). A connection between the minimization of the sum in Equation (4.8) and the ML theory is provided by ALD (Yu & Zhang, 2005). ALD that is closely associated with the loss function for QR has been examined in several works of literature (Yu & Zhang, 2005; Yu & Moyeed, 2001; Geraci & Bottai, 2007; Liu & Bottai, 2009; Lachos et al., 2015; Kotz et al., 2002). Other forms of Laplace distribution were summarized by Kozubowski & Nadarajah (2010). A book entirely devoted to the Laplace distribution, historical background, and its extension is presented by Kotz et al. (2012).

As discussed in Koenker & Machado (1999) and Yu & Moyeed (2001) that a continuous r.v. $Y \in \mathbb{R}$ is distributed as an ALD if its probability density function (pdf) with location parameter μ, scale parameter $\sigma > 0$, skewness parameter $\tau \in (0,1)$, and $\rho_\tau(u) = u(\tau - I\{u < 0\})$ represents the contribution by residuals u, is given by

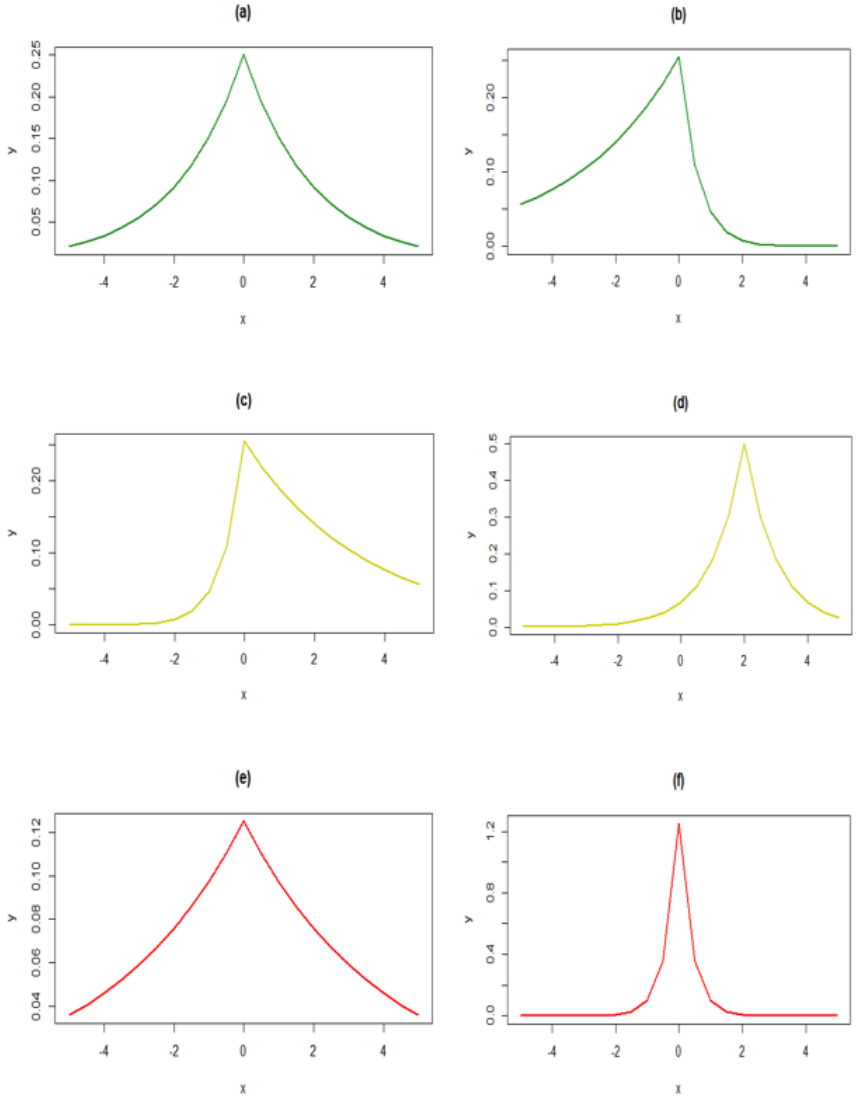
$$f(Y|\mu,\sigma,\tau) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{-\rho_\tau\left(\frac{y_i - \mu_i}{\sigma}\right)\right\}, \tag{4.10}$$

which can be denoted as $y \sim ALD(\mu,\sigma,\tau)$, then $Pr(y \leq \mu) = \tau$ and $Pr(y > \mu) = 1-\tau$ indicates that the parameters μ and τ in ALD satisfy μ to be the τ^{th} quantile of y. QR adapt this important feature of ALD (Yu et al., 2003). Moreover, Yu & Zhang (2005) addressed detail investigation of the various properties and generalization of an ALD.

Since $\tau \in (0,1)$ is the skewness parameter, the ALD splits along the scale parameter into two parts, one with probability τ to the left and one with probability $(1-\tau)$ to the right. That is, ALD(μ,σ,τ) is negatively skewed when $\tau > 0.5$, and positively skewed when $\tau < 0.5$. When μ increases, the density shifted on the x-axis. For higher σ, the density is wider, and the data is more spreader. Figure 4.1 displays the ALD densities of these cases for a random sample generated from R; however, see Yu & Zhang (2005) for more graphical representation of an ALD density. The three-parameter ALD defined in Koenker & Machado (1999) useful for QR is implemented in the R package '*ald*'. It provides the probability density function, distribution function, quantile function, random number generator function, likelihood function, moments, and ML estimator for a given sample (Galarza & Galarza, 2015). Further, the ALD, which is characterized by three parameters μ, σ, and τ (Equation (4.10)) reduces into two special cases:

$$f(Y|\mu,\sigma,\tau) = \begin{cases} \tau(1-\tau)\exp\left\{-\rho_\tau(y-\mu)\right\}, & \text{if} \quad \sigma = 1 \\ \frac{1}{4\sigma}\exp\left\{-\frac{|y-\mu|}{2\sigma}\right\}, & \text{if} \quad \tau = 0.5, \end{cases}$$

where the first case is considered by Koenker & Machado (1999), and the second case is usually called *symmetric* (double-exponential) Laplace distribution with location parameter μ and scale parameter 2σ (Yu & Zhang, 2005). Several of the extended ALDs are based on either the mixture of the *symmetric* Laplace distribution or a split of it (Kotz et al., 2002; Kozubowski & Nadarajah, 2010). See also Kotz et al. (2012) or Kozumi & Kobayashi (2011) for more details, and representation of the various mixture of an ALD in the ALD(μ,σ,τ) parameterization.



(a) for ALD (0, **1**, 0.5); (b) for ALD (0, 0.5, **0.85**); (c) for ALD (0, 0.5, **0.15**); (d) for ALD (**2**, 0.5, 0.5); (e) for ALD (0, **2**, 0.5); (f) for ALD (0, **0.2**, 0.5).

Figure 4.1: Densities of an Asymmetric Laplace Distribution

As illustrated in Yu & Zhang (2005), the k^{th} central moment of an ALD for a r.v. $Y \sim ALD(\mu, \sigma, \tau)$ is written as follows

- The k^{th} central moment

$$E(Y-\mu)^k = k!\sigma^k\tau(1-\tau)\left(\frac{1}{\tau^{k+1}} + \frac{(-1)^k}{(1-\tau)^{k+1}}\right) \tag{4.11}$$

Thus, the mean and variance of an ALD can be derived from this moment:

- The mean of an ALD can be derived as follows

$$\begin{aligned}
E(Y) &= E(Y-\mu+\mu), \qquad \text{where} \quad \mu \quad \text{is constant} \\
&= E(Y-\mu)^1 + E(\mu), \qquad k=1 \\
&= \sigma\tau(1-\tau)\left(\frac{1}{\tau^2} + \frac{-1}{(1-\tau)^2}\right) + \mu, \qquad \text{using} \quad \text{Equation(4.11)} \\
&= \sigma\tau(1-\tau)\left(\frac{(1-\tau)^2-\tau^2}{\tau^2(1-\tau)^2}\right) + \mu \\
&= \frac{\sigma(1-2\tau+\tau^2-\tau^2)}{\tau(1-\tau)} + \mu \\
&= \mu + \frac{\sigma(1-2\tau)}{\tau(1-\tau)}
\end{aligned} \tag{4.12}$$

- The variance of an ALD can also be derived as follows

$$\begin{aligned}
Var(Y) &= Var(Y-\mu+\mu), \qquad \text{where} \quad \mu \quad \text{is constant} \\
&= Var(Y-\mu) + Var(\mu) \\
&= Var(Y-\mu), \quad Var(\mu) = 0 \quad \text{by properties of variance} \\
&= E((Y-\mu)^2) - (E(Y-\mu))^2, \quad \text{when} \quad k=2 \quad \text{in Equation(4.11)} \\
&= 2\sigma^2\tau(1-\tau)\left(\frac{1}{\tau^3} + \frac{1}{(1-\tau)^3}\right) - \left(\sigma\tau(1-\tau)\left(\frac{1}{\tau^2} + \frac{-1}{(1-\tau)^2}\right)\right)^2 \\
&= 2\sigma^2\tau(1-\tau)\left(\frac{(1-\tau)^3+\tau^3}{\tau^3(1-\tau)^3}\right) - \left(\frac{\sigma(1-2\tau)}{\tau(1-\tau)}\right)^2 \\
&= \frac{2\sigma^2(1-3\tau+3\tau^2)}{\tau^2(1-\tau)^2} - \frac{\sigma^2(1-4\tau+4\tau^2)}{\tau^2(1-\tau)^2} \\
&= \frac{\sigma^2(2-6\tau+6\tau^2-1+4\tau-4\tau^2)}{\tau^2(1-\tau)^2} \\
&= \frac{\sigma^2(1-2\tau+2\tau^2)}{\tau^2(1-\tau)^2}
\end{aligned} \tag{4.13}$$

For independently distributed r.v. $Y_i|x_i \overset{i}{\sim} \text{ALD}(\mu_i, \sigma, \tau)$ with $\mu_i = \boldsymbol{x}_i'\boldsymbol{\beta}_\tau$, the likeli-

hood density function of an n-dimensional ALD is given as

$$\begin{aligned} L(\beta, \sigma | y, \tau) &= \prod_{i=1}^{n} f(Y | \mu_i, \sigma, \tau) \\ &= \prod_{i=1}^{n} \frac{\tau(1-\tau)}{\sigma} \exp\left\{-\rho_\tau\left(\frac{y_i - \mu_i}{\sigma}\right)\right\} \\ &= \prod_{i=1}^{n} \frac{\tau(1-\tau)}{\sigma} \exp\left\{-\rho_\tau\left(\frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_\tau}{\sigma}\right)\right\} \\ &= \frac{\tau^n(1-\tau)^n}{\sigma^n} \exp\left\{-\sum_{i=1}^{n} \rho_\tau\left(\frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_\tau}{\sigma}\right)\right\}, \end{aligned} \tag{4.14}$$

for a fixed $\tau \in (0,1)$, Equation (4.14) is proportional to

$$L(\beta, \sigma | y, \tau) \propto \sigma^{-n} \exp\left\{-\sum_{i=1}^{n} \rho_\tau\left(\frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_\tau}{\sigma}\right)\right\}.$$

Thus, the ML estimator of μ_i is given by $\hat{\mu}_i = \boldsymbol{x}_i'\boldsymbol{\beta}_\tau$ with

$$\hat{\beta}_\tau = \underset{\beta \in \mathbb{R}^p}{\arg\max} \left\{\sigma^{-n} \exp\left(-\sum_{i=1}^{n} \rho_\tau\left(\frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_\tau}{\sigma}\right)\right)\right\} \tag{4.15}$$

This shows that for a fixed $\tau \in (0,1)$ the estimators $\hat{\beta}_\tau$ from Equation (4.8) and from Equation (4.15) align; it holds that

$$\begin{aligned} \hat{\beta}_\tau &= \underset{\beta \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_\tau) \\ &= \underset{\beta \in \mathbb{R}^p}{\arg\max} \left\{\sigma^{-n} \exp\left(-\sum_{i=1}^{n} \rho_\tau\left(\frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_\tau}{\sigma}\right)\right)\right\}, \end{aligned}$$

which implies that maximizing the likelihood function in Equation (4.15) with respect to β is equivalent to minimizing the objective function in Equation (4.8) (Liu & Bottai, 2009; Lachos et al., 2015). This matches the result from simple linear regression, where the OLS estimator of the regression parameter minimizing the error sum of squares is equivalent to the ML estimator of the corresponding Gaussian likelihood (Galarza et al., 2020).

4.2.4 Quantile Regression for Count Data

The conventional QR is based on the median, or other quantile levels, by assuming a continuous or Gaussian distribution. QR has been extended to count regression, which is a special case of the discrete variable model (Winkelmann, 2008; Hilbe, 2011, 2014; Machado & Silva, 2005; Cameron et al., 2009; Cameron & Trivedi, 2013). However, the distribution function of a discrete r.v. is not continuous, and the objective of interest of the conditional quantile $Q_\tau(y|\boldsymbol{x})$ for discrete distribution cannot be a continuous function of $\boldsymbol{x}$ such as $exp(\boldsymbol{x}'\beta)$ (Winkelmann, 2008). Machado & Silva (2005) overcome this restriction by developing a continuous r.v. whose quantiles have a one-to-one relation with the quantiles of y. The key step in their model is to replace the count response, y, with a continuous r.v., $z = h(y)$, where $h(\cdot)$ is a smooth continuous transformation. Hence, the count response, y, is structured as

$$z = y + u, \tag{4.16}$$

where u $\sim$uniform$(0, 1)$ is a *pseudorandom* draw from the uniform distribution on $(0, 1)$. This step is also known as "jittering", which is the process of eliminating the discontinuities in the Poisson or negative binomial count models or any other models for count data in such a way that the resultant distribution appears as a continuous variable; thus, the entire conditional quantile can then be model (Hilbe, 2011, 2014; Cameron et al., 2009; Cameron & Trivedi, 2013; Machado & Silva, 2005). Stata command: *qcount* is available to model QR for count data using the *jittering* method (Hilbe, 2011; Miranda, 2007; Cameron et al., 2009).

As illustrated in Machado & Silva (2005), the following parameterization is used to represent a transformation, $T(z;\tau)$, and its associated representation of the conditional τ-quantile of z, $Q_\tau(z|\boldsymbol{x})$:

$$T(z;\tau) = \begin{cases} \log(\xi), & z \leq \tau \\ \log(z-\tau), & z > \tau, \qquad \tau \in (0,1) \end{cases}$$

where $0 < \xi < \tau$, which is a small positive number, and $\boldsymbol{x}$ represents a vector of explanatory variables. Hence, it follows that the transformed *jittered* quantile function written as

$$Q_\tau(T(z;\tau)|\boldsymbol{x}) = \boldsymbol{x}'\beta_\tau$$

Winkelmann (2006, 2008) stated that after $Q_\tau(z|\boldsymbol{x})$ and $T(\cdot)$ are specified, the parameters, $\boldsymbol{\beta}_\tau$, estimated as solution to

$$\min \sum_{i=1}^{n} \rho_\tau(T(z;\tau) - \boldsymbol{x}'\beta_\tau)$$

where $\rho_\tau(u) = u(\tau - I\{u < 0\})$.

From Equation (4.16), one can study the conditional quantiles of z, $Q_\tau(z|\boldsymbol{x})$, but it is τ^{th} quantiles of the conditional distribution of y, $Q_\tau(y|\boldsymbol{x})$, that are of interest (Cameron et al., 2009).

The conditional quantile for $Q_\tau(z|\boldsymbol{x})$ is specified as

$$Q_\tau(z|\boldsymbol{x}) = \tau + exp(\boldsymbol{x}'\beta), \qquad 0 < \tau < 1 \tag{4.17}$$

The conditional quantile function of the objective interest, $Q_\tau(y|\boldsymbol{x})$ is

$$Q_\tau(y|\boldsymbol{x}) = \lceil Q_\tau(z|\boldsymbol{x}) - 1 \rceil, \tag{4.18}$$

where $\lceil a \rceil$ denotes the *ceiling function* that returns the smallest integer greater than, or equal to, a (Machado & Silva, 2005).

When the count data consists of severe outliers or multiple distributional components that do not reflect a known underlying probability distribution, quantile count models may be a useful alternative. Furthermore, QR models all of the quantiles of distribution and covers the entire range of counts (Hilbe, 2011). For this reason, QR to *jittered* count data can be used as a valuable additional tool to make inferences about the entire range of counts, but it cannot replace the more structured and well-proven models for count data analysis (Machado & Silva, 2005). Hilbe (2011, 2014) also stated that this new and growing in use class of count model, quantile count model, could be used when the distribution, or mixture of distributions, cannot be identified. Detailed discussions about quantile count models for independent data are available in Winkelmann (2008), Machado & Silva (2005), Hilbe (2011, 2014), Cameron et al. (2009), Cameron & Trivedi (2013), and the recent application of this model can be found in Winkelmann (2006) and Miranda (2008).

4.3 Quantile Mixed-Effects Models

Although QR was at first developed under a univariate system, the considerable amount of longitudinal data recently produces its extensions toward mixed-effects modeling system through either the distribution-free way (Galvao Jr, 2011; Fu & Wang, 2012; Lipsitz et al., 1997) or the likelihood-based way in most cases following the ALD (Galarza et al., 2015; Geraci & Bottai, 2007, 2014; Galarza et al., 2017). The likelihood-based quantile mixed model additionally makes use of different parametric distributions such as an infinite mixture of Gaussian densities (Reich et al., 2010), and direct parametric maximum likelihood (ML) approach (Noufaily & Jones, 2013). The distribution-free approaches that consist of fixed-effects and weighted generalize estimating equations consider the use of independent estimating equations that ignore correlations between repeated measurements leads to loss of efficiency (Geraci & Bottai, 2014; Koenker, 2004; Lipsitz et al., 1997). Meanwhile, Geraci & Bottai (2007) suggested a likelihood-based QR model for longitudinal data that accounts for within-subject dependence by incorporating subject-level random effects and modeling the residual distribution with an ALD. Liu & Bottai (2009) also developed a likelihood-based method to estimate parameters of conditional quantile functions with random effects by incorporating an ALD for the random error term that is not restricted to be normal. The within-subject correlation among data is taken into account by adding random effects to get unbiased parameter estimates (Liu & Bottai, 2009).

Although the application of QR for mixed-effects models has received increasing consideration in wide-ranging areas of study, such as marine biology, environmental science, cardiovascular disease, and ophthalmology (Geraci & Bottai, 2007; Muir et al., 2015; Fornaroli et al., 2015; Blankenberg et al., 2016; Patel et al., 2016); it has not easily extended to the mixed-effects models because of the greater complexity in solving the minimization problem.

Mixed-effects models characterize an ordinary and conventional type of regression methods used to examine data coming from longitudinal studies. Recall the general linear mixed-effects model:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \boldsymbol{Z}_i'\boldsymbol{u}_i + \varepsilon_{ij}, \qquad i = 1, ..., n, \qquad j = 1, ..., n_i,$$

where $\boldsymbol{Y}_i$ is the $n_i \times 1$ vector of response variable, $\boldsymbol{X}_i'$ is the i^{th} row of a known $n_i \times p + 1$ design matrix, $\boldsymbol{\beta}$ is $p \times 1$ vector of population-averaged fixed-effects, $\boldsymbol{Z}_i$ with the dimension of $n_i \times q + 1$ known design matrix for random effects, $\boldsymbol{u}_i$ is $q \times 1$

vector of random effects, $\boldsymbol{u}_i \sim N(0, \Sigma_u)$, and $\varepsilon_{ij} \sim N(0, \sigma^2)$. As previously discussed, Koenker & Bassett (1978) defined the quantile regression with no random effects (Equation (4.7)), in which their model considered observations are independent. Koenker (2004) introduced an approach of quantile regression with fixed effects models for the application of longitudinal data. He considered the conditional quantile functions of the response of the j^{th} observation on the i^{th} individual y_{ij} of the form

$$Q_\tau(y_{ij}|\boldsymbol{x}_{ij}) = \boldsymbol{\alpha}_i + \boldsymbol{x}'_{ij}\boldsymbol{\beta}_\tau + \varepsilon_{ij}, \qquad j = 1, ..., n_i; \quad i = 1, ..., M \tag{4.19}$$

By solving,

$$\arg\min_{(\alpha,\beta)} \sum_{k=1}^{q} \sum_{j=1}^{n_i} \sum_{i=1}^{M} w_k \rho_{(\tau k)}(y_{ij} - \boldsymbol{\alpha}_i - \boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(\tau k)}), \tag{4.20}$$

to estimate the model for several quantiles simultaneously. Note that $\rho_\tau = u(\tau - I(u < 0))$ indicates the piecewise linear quantile loss function of Koenker & Bassett (1978), w_k denotes the weights that control the relative influence of the q quantiles $\{\tau_1, ..., \tau_q\}$, on the estimation of α_i parameters, and α_i's have a pure location shift effect on the conditional quantiles of the response, which were added to capture some individual specific source of variability (unobserved heterogeneity) that was not adequately controlled for by other covariates in the model. The effect of the covariates, $\boldsymbol{x}_{ij}$ are allowed to depend upon the quantile, τ, of interest, but the α_i's do not (Koenker, 2004). However, this approach does not account for between/within subjects' variability; thus, it is computationally inefficient when dealing with large sample sizes or complex models.

Furthermore, Koenker (2004) also considered using penalized quantile regression with subject-specific fixed-effects to model longitudinal data. In this model, Konker considered estimators of the penalized version of Equation (4.15) by solving

$$\arg\min_{(\alpha,\beta)} \sum_{k=1}^{q} \sum_{j=1}^{n_i} \sum_{i=1}^{M} w_k \rho_{(\tau k)}(y_{ij} - \boldsymbol{\alpha}_i - \boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(\tau k)}) + \lambda \sum_{i=1}^{n} |\alpha_i|, \tag{4.21}$$

which considers the L_1-penalty, $P(\alpha) = \sum_{i=1}^{n} |\alpha_i|$ for the loss function ρ_τ, rather than the conventional Gaussian penalty, to maintain the LP form of the problem and also to preserve the sparsity of the resulting design matrix (Koenker, 2004). Equation (4.17) reduced to the following form of a quantile function by Koenker (2004) for each individual

$$Q_\tau(y_{ij}|\boldsymbol{x}_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(\tau)} + \alpha_{i(\tau)} \tag{4.22}$$

As a result, Koenker (2004) obtained two component parts: a common population quantile function and an individual component for each subject. As Konker suggested, when the population is of interest, a penalty method that controls for the subject variability can be a useful approach due to the computational simplicity (Koenker, 2004).

Similarly to the conventional QR model (Equation (4.7)) without random effect for a fixed $\tau \in (0,1)$, Geraci & Bottai (2007) extended the QR model by incorporating random intercept to model longitudinal data that follows ALD for the conditional response. Thus, the dependence among data (within-group correlation) accounted for by a subject-specific random intercept, modeled as

$$Q_\tau(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{u}_{0i}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta}_\tau + \boldsymbol{u}_{0i} + \varepsilon_{\tau,ij}, \qquad \text{with} \quad Q_{\varepsilon_{\tau,ij}}(\tau|\boldsymbol{x}_{ij}, \boldsymbol{u}_{0i}) = 0 \tag{4.23}$$

where the extended random effect $\varepsilon_{\tau,ij}$ distributes as $\varepsilon_{\tau,ij} \overset{iid}{\sim} ALD(0, \sigma, \tau)$ that carries τ in the footnote, which implies for different quantile level the random effect may be different. The ALD, which was discussed in subsection (4.2.3), also serves as the distribution of the individual error term $\varepsilon_{\tau,ij}$ here. As Geraci & Bottai (2007) presented, estimation of the regression quantiles for this approach was then accomplished using Gibbs Sampling, which is also highlighted in Appendix C.

The quantile mixed model version of Equation (4.19) only needs to be employed whenever the distribution of the error term in the mixed model (Equation (4.19)) is unknown. But for a known error distribution function F_ε, which is not practical in real data application, the τ-quantile of y_{ij} given $\boldsymbol{x}_{ij}$ would be

$$Q_\tau(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{u}_{0i}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta}_\tau + \boldsymbol{u}_{0i} + \varepsilon_{\tau,ij} + F^{-1}_{\varepsilon(\tau)}, \tag{4.24}$$

where $i = 1, 2, ..., M; j = 1, 2, ..., n_i$, and $\boldsymbol{\beta}$ is the same parameter vector as in Equation (4.19). However, assuming an unknown error distribution ε for practical application leads to model flexibility.

Linear QR with multiple random effects (random intercept and random slope) simply general linear quantile mixed-effects models (QR-LMM hereafter) was developed by Geraci & Bottai (2014) as an extension of their previous work of Geraci & Bottai (2007). The τ^{th} QR-LMM is modeled as

$$Q_\tau(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{u}_i) = \boldsymbol{x}'_{ij}\boldsymbol{\beta}_\tau + \boldsymbol{z}'_{ij}\boldsymbol{u}_i + \varepsilon_{\tau,ij}, \qquad \text{with} \quad Q_{\varepsilon_{\tau,ij}}(\tau|\boldsymbol{x}_{ij}, \boldsymbol{u}_i) = 0 \tag{4.25}$$

where the random errors $\varepsilon_{\tau,ij} \sim ALD(0,\sigma,\tau)$ are also dependent on τ, $\boldsymbol{\beta}_\tau$ is the coefficient of fixed-effects corresponding to the τ^{th} quantile, and the continuous response variable y_{ij}, conditional on $\boldsymbol{x}_{ij}$ and $\boldsymbol{u}_i$ for $i = 1, ..., n$, $j = 1, ..., n_i$ are assumed to be independently distributed as ALD with the density given by

$$f(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{u}_i, \sigma) = \frac{\tau(1-\tau)}{\sigma} exp\left\{-\rho_\tau\left(\frac{y_{ij} - \boldsymbol{x}'_{ij}\boldsymbol{\beta}_\tau - \boldsymbol{z}'_{ij}\boldsymbol{u}_i}{\sigma}\right)\right\}, \tag{4.26}$$

The random effects ($\boldsymbol{u}'_i s$) are assumed to be distributed as $u_i \overset{iid}{\sim} N_\tau(0, \boldsymbol{\Psi})$, where the dispersion matrix $\boldsymbol{\Psi} = \boldsymbol{\Psi}(\boldsymbol{\alpha})$ relies on unknown and reduced parameters $\boldsymbol{\alpha}$ (Lachos et al., 2015; Galarza, 2015). Then a likelihood for y_{ij} at the τ^{th} quantile is

$$L(\boldsymbol{\beta}, \sigma|y_{ij}, \tau) = \frac{\tau^n(1-\tau)^n}{\sigma^n} exp\left\{-\sum_{i=1}^{n}\sum_{j=1}^{n_i}\rho_\tau\left(\frac{y_{ij} - \boldsymbol{x}'_{ij}\boldsymbol{\beta}_\tau - \boldsymbol{z}'_{ij}\boldsymbol{u}_i}{\sigma}\right)\right\} \tag{4.27}$$

Based on the likelihood of conditional quantile of y_{ij}, it is suggested that the maximization of the likelihood in Equation (4.26) with respect to the parameter β_τ is equivalent to the minimization of the loss function in Equation (4.28). Thus, we can estimate the coefficient of fixed-effects corresponding to the τ^{th} quantile ($\boldsymbol{\beta}_\tau$) by minimizing the objective function of Equation (4.27), which can be expressed as

$$H^*(\boldsymbol{\beta}_\tau) = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\rho_\tau\left(\frac{y_{ij} - \boldsymbol{x}'_{ij}\boldsymbol{\beta}_\tau - \boldsymbol{z}'_{ij}\boldsymbol{u}_i}{\sigma}\right) \tag{4.28}$$

As described by Geraci & Bottai (2007, 2014), the QR-LMM need an estimator (maximum likelihood) for the parameter β_τ, and a predictor for the random vector leading to the conditional quantile function estimator for a fixed $\tau \in (0, 1)$.

$$\hat{Q}_\tau(y_{ij}|\boldsymbol{x}_{ij}) = \boldsymbol{x}_{ij}\hat{\boldsymbol{\beta}}_\tau + \hat{\varepsilon}_{\tau,ij}$$

These two estimation process: *maximum likelihood methods* and *prediction of random effects* are summarized herein, which are also described in Geraci & Bottai (2007, 2014). The methods are implemented in the open software R (see the package *lqmm* by Geraci et al. (2014)).

Maximum Likelihood Methods

From the QR-LMM (Equation (4.19)), the conditional distribution of $F_{y_{ij}|\boldsymbol{u}_i}$ is assumed to be unknown, and follows an ALD, with location, scale, and skewness parameters given $\mu_{\tau,ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta}_\tau + \boldsymbol{z}'_{ij}\boldsymbol{u}_i$, σ_τ, and τ, respectively, where $\boldsymbol{\beta}_\tau \in \mathbb{R}^p$ is

a vector of unknown fixed effects, and τ defines the quantile level to be estimated (Geraci et al., 2014).

Thus the joint desity of the observation vector $\boldsymbol{y}$ and the random effect vector $\boldsymbol{u}$, $(\boldsymbol{y}, \boldsymbol{u})$, for QR-LMM is given as

$$\begin{aligned} f\left(\boldsymbol{y}, \boldsymbol{u} | \boldsymbol{\beta}_\tau, \sigma_\tau, \boldsymbol{\Psi}_\tau\right) &= f\left(\boldsymbol{y} | \boldsymbol{\beta}_\tau, \sigma_\tau, \boldsymbol{u}\right) f\left(\boldsymbol{u} | \boldsymbol{\Psi}_\tau\right) \\ &= \prod_{i=1}^{M} f\left(\boldsymbol{y}_i | \boldsymbol{\beta}_\tau, \sigma_\tau\right) f\left(\boldsymbol{u}_i | \boldsymbol{\Psi}_\tau\right) \\ &= \prod_{i=1}^{M} \left(\prod_{j=1}^{n_i} f\left(\mu_{\tau,ij} | \boldsymbol{\beta}_\tau, \sigma_\tau\right) f\left(\boldsymbol{u}_i | \boldsymbol{\Psi}_\tau\right) \right) \end{aligned} \tag{4.29}$$

where $\boldsymbol{u}_i = (u_{i1}, ..., u_{iq})'$, for $i = 1, ..., M; j = 1, ..., n_i$, assumed to be a zero-median random vector independent from the model's error term and distributed according to $f\left(\boldsymbol{u}_i | \boldsymbol{\Psi}_\tau\right)$, and $\boldsymbol{\Psi}_\tau$ is a $q \times q$ covariance matrix (Geraci et al., 2014).

By integrating out the random effects $\boldsymbol{u}$ from Equation (4.25), the marginal likelihood can be obtained:

$$\begin{aligned} L(\boldsymbol{y} | \boldsymbol{\Psi}_\tau, \boldsymbol{\beta}_\tau, \sigma_\tau) &= \int_{\mathbb{R}^q} \prod_{i=1}^{M} \left(\prod_{j=1}^{n_i} f\left(\mu_{\tau,ij} | \boldsymbol{\beta}_\tau, \sigma_\tau\right) f\left(\boldsymbol{u}_i | \boldsymbol{\Psi}_\tau\right) \right) du_i \\ &= \prod_{i=1}^{M} \int_{\mathbb{R}^q} \left(\prod_{j=1}^{n_i} f\left(\mu_{\tau,ij} | \boldsymbol{\beta}_\tau, \sigma_\tau\right) f\left(\boldsymbol{u}_i | \boldsymbol{\Psi}_\tau\right) \right) du_i \\ &= \prod_{i=1}^{M} \int_{\mathbb{R}^q} \left(\prod_{j=1}^{n_i} f\left(\boldsymbol{x}'_{ij} \boldsymbol{\beta}_\tau + \boldsymbol{z}'_{ij} \boldsymbol{u}_i | \boldsymbol{\beta}_\tau, \sigma_\tau\right) f\left(\boldsymbol{u}_i | \boldsymbol{\Psi}_\tau\right) \right) du_i, \end{aligned} \tag{4.30}$$

where $\mathbb{R}^q$ denotes the q-dimensional Euclidean space (Geraci & Bottai, 2014).

Since there is no analytical or closed-form solution for the above integral, approximation methods such as marginal, MCMC methods, and numerical integration are needed (Geraci & Bottai, 2007, 2014). Geraci & Bottai (2007) first attempt to estimate the parameter of the QR model with random intercept by using a Monte Carlo EM (MCEM) algorithm, which, however, found to be computationally intensive and inefficient (Geraci & Bottai, 2014). Later, Geraci & Bottai (2014) made use of a different approach based on a Gaussian quadrature, which is also implemented in the R package *lqmm* (Geraci et al., 2014), shows advanced than the previous MCEM method.

By using numerical integration technique, Geraci & Bottai (2014) derived the marginal log-likelihood density, denoted as $\ell(\boldsymbol{y}|\boldsymbol{\Psi}_\tau, \boldsymbol{\beta}_\tau, \sigma_\tau) = \log L(\boldsymbol{y}|\boldsymbol{\Psi}_\tau, \boldsymbol{\beta}_\tau, \sigma_\tau)$, written as

$$\ell = \sum_{i=1}^{M} \left[\log \left(\frac{\tau^{n_i}(1-\tau)^{n_i}}{\sigma^{n_i}} \right) + \log \int_{\mathbb{R}^q} \exp \left\{ -\frac{1}{\sigma} \sum_{j=1}^{n_i} \rho_\tau(y_{ij} - \mu_{\tau,ij}) \right\} f\left(\boldsymbol{u}_i|\boldsymbol{\Psi}_\tau\right) du_i \right], \tag{4.31}$$

where, the random effect $\boldsymbol{u}$ is assumed to be normally distributed as $\boldsymbol{u} \sim N(0, \boldsymbol{\Psi})$, which leads to a Gauss-Hermite quadrature for the **approximate** ALD-based log-likelihood, denoted as $\ell_{app}(\boldsymbol{y}|\boldsymbol{\Psi}_\tau, \boldsymbol{\beta}_\tau, \sigma_\tau)$, see Geraci et al. (2014), or Geraci & Bottai (2014), for further discussion.

Prediction of Random Effects

The predictor of the random effects $\boldsymbol{U}$ for the τ^{th} QR-LMM can be written as

$$\hat{\boldsymbol{U}}_\tau = \hat{\boldsymbol{\Psi}}_\tau \boldsymbol{Z}_\tau \hat{\boldsymbol{\Sigma}}^{-1} \left\{ \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\tau - \hat{E}[\Sigma_\tau] \right\}, \tag{4.32}$$

where the estimated covariance matrix of $\boldsymbol{Y}$, which is $\boldsymbol{\Sigma} = \boldsymbol{Z}\hat{\boldsymbol{\Psi}}_\tau \boldsymbol{Z}' + \hat{Var}(\Sigma_\tau)$, and the estimated mean and variance of the ALD with parameters $\mu = 0, \hat{\sigma}$, and τ that are given in Yu & Zhang (2005) can also be written here as

$$\begin{aligned} \hat{E}[\Sigma_\tau] &= \frac{\hat{\sigma}(1-2\tau)}{\tau(1-\tau)} \\ \hat{Var}(\Sigma_\tau) &= \frac{\hat{\sigma}^2(1-2\tau+2\tau^2)}{\tau^2(1-\tau)^2} \end{aligned} \tag{4.33}$$

Note that Geraci & Bottai (2014) used an approach based on the best linear predictor (BLP) of Ruppert et al. (2003) for prediction of $\boldsymbol{U}$. As a result QR-LMM estimator, $\hat{Q}_\tau(y_{ij}|\boldsymbol{x}_{ij})$, is a combination of $\hat{\beta}_\tau$ (maximum likelihood estimator) from the first estimation process, and $\hat{U}_\tau$ (predictor of random effect) given in Equation (4.27). However, Geraci & Bottai (2014) stated that prediction of random effects in QR-LMMs is still an ongoing research issue. More details regarding the estimation process of quantile mixed-effects models are available here (Geraci & Bottai, 2007, 2014; Yu & Zhang, 2005; Galarza et al., 2015; Liu & Bottai, 2009; Geraci et al., 2014).

Since computational issues for longitudinal quantile regression are still an open problem, different approaches have been investigated by several researchers (Marino & Farcomeni, 2015). Recently, Galarza et al. (2015, 2017) presented a robust parametric ALD-based QR-LMM that follows the stochastic approximation of the expectation-maximization (SAEM) algorithm for deriving **exact** ML estimates of the fixed-effects, and the general variance-covariance matrix $\Sigma_\tau = \Sigma(\boldsymbol{\theta}_\tau)$ of the random effects pa-

rameters for the specific quantile. The SAEM estimating algorithm for QR-LMM is implemented in the open software R, see the package *qrLMM* by Galarza et al. (2017).

4.3.1 The EM and SAEM algorithms

The Expectation-Maximization algorithm, also known as the EM algorithm, which was proposed by Dempster et al. (1977), is a popular technique to the iterative computation of ML estimates when the observations can be viewed as incomplete data, which incorporates the ordinary sense of missing data; however, it is much broader than that (McLachlan & Krishnan, 2007). There are two steps in each iteration of the EM algorithm: an expectation, or E-step, followed by a maximization (M-step). "In the former action, the incomplete data are estimated given the observed data and current estimates of the model parameters. In the later step, the likelihood function is maximized under the assumption that the incomplete/missing data is known" (Dempster et al., 1977). The detailed explanations of these processes, their related analytical clarifications for successively more common sorts of models, and the basic theory underlying the EM algorithm are presented in Dempster et al. (1977). A book devoted entirely to the general formulation of the EM algorithm, as well as its basic properties and applications, has been provided by McLachlan & Krishnan (2007). Moreover, the success of the EM algorithm is well documented and can be found in numerous statistical literature.

Even though the EM algorithm is popular, Delyon et al. (1999) pointed out that, in certain circumstances, it is not applicable due to the fact that the E-step cannot be carried out in a closed-form. To bargain with these issues, Delyon et al. (1999) presented a simulation-based SAEM algorithm as an elective to the MCEM, standing for Monte Carlo EM. "While the MCEM requires a consistent increment of the simulated data and regularly a substantial amount of simulations, the SAEM versions guarantee convergence with a fixed and/or small simulation size" (Meza et al., 2012; Delyon et al., 1999; Jank, 2006). The SAEM algorithm replaces the E-step of the EM algorithm by one iteration of a stochastic (probabilistic) approximation procedure, whereas the M-step is consistent (Meza et al., 2012). The E- and M-steps of the EM and SAEM procedures are highlighted herein. For more points of interest, however, see Jank (2006), Meza et al. (2012), or Kuhn & Lavielle (2004, 2005). Furthermore, details of these algorithms for estimating the parameters of the QR-LMM are presented by (Galarza et al., 2015, 2017). "The SAEM algorithm was proven to be more effective for computing the ML estimates in mixed-effects models due to the reusing of simulations from one iteration to the next in the smoothing phase of the algorithm" (Meza et al., 2012; Kuhn & Lavielle, 2004, 2005; Galarza et al., 2015). The SAEM al-

gorithm is implemented in the R package *qrLMM*().

Let $\ell_o(\hat{\theta}) = logf(Y_{obs};\theta)$ denotes the maximization of log-likelihood function based on the observed data (Y_{obs}), q represents missing data, $Y_{com} = (Y_{obs}, q)'$ denotes the complete data with observed and missing data, $\ell_c(Y_{com};\theta)$ be the complete log-likelihood function, and $\hat{\theta}_k$ indicates the estimated value of θ at the k^{th} iteration. Then the EM algorithm in modeling with missing data, that maximizes $\ell_c(Y_{com};\theta) = \log f(Y_{obs}, q;\theta)$ iteratively and converges to a stationary point of the observed likelihood under mild regularity conditions (Meza et al., 2012; Galarza et al., 2015), go through in two steps:

- **E-step**: Consists computing of the conditional expectation of $\ell_c(Y_{com};\theta)$

$$S(\theta|\hat{\theta}_k) = E\left\{\ell_c(Y_{com};\theta)|Y_{obs},\hat{\theta}_k\right\} \tag{4.34}$$

- **M-step**: Computes the parameter values $\hat{\theta}_{k+1}$ as maximizing $S(\theta|\hat{\theta}_k)$ with respect to θ.

The SAEM algorithm, which replaces the E-step by stochastic approximation, presented by Galarza et al. (2015), is summarized as follows:

- **Simulation (E-step)**: Generate $q(\ell_o, k)$ sample (simulation of the missing data at iteration k), $\ell = 1, 2, ..., m$, from the conditional distribution of the missing data $f(q|\theta_{k-1}, Y_{obs})$
- **Stochastic approximation**: Update $S(\theta|\hat{\theta}_k)$ according to

$$S(\theta|\hat{\theta}_k) = S(\theta|\hat{\theta}_{k-1}) + \delta_k\left[\frac{1}{m}\sum_{\ell=1}^{m}\ell_c(Y_{obs}, q(\ell_o, k)|\hat{\theta}_k;\theta) - S(\theta|\hat{\theta}_{k-1})\right] \tag{4.35}$$

- **M-step**: Maximize $\hat{\theta}_k$ according to

$$\hat{\theta}_{k+1} = \arg\max_{\theta} S(\theta|\hat{\theta}_k),$$

this is equivalent to finding $\hat{\theta}_{k+1} \in \Theta$ such that $S(\hat{\theta}_{k+1}) \geq S(\hat{\theta}_k)$,

where δ_k is a smoothing parameter (a sequence of decreasing non-negative numbers) as given by Kuhn & Lavielle (2004), and m is the number of simulations suggested to be less than or equal to 20 (Galarza et al., 2015). The choice of δ_k recommended by

Galarza et al. (2015) is given as follows:

$$\delta_k = \begin{cases} 1 & \text{for} \quad 1 \leq k \leq cW \\ \frac{1}{k-cW} & \text{for} \quad cW+1 \leq k \leq W, \end{cases}$$

where $c \in (0,1)$ is a cut point that regulates the percentage of initial iterations with no memory, and W is the maximum number of iterations.

4.3.2 Quantile Regression for Longitudinal Count Data

As discussed in Subsection (4.2.4), the quantiles of count data must be integers due to the fact that counts themselves are integers. Since the QR-LMM (Equation (4.21)) is a model for continuous data, it is not directly applicable on counts. To date, the application of QR for examining longitudinal count data is not fully developed. However, the application of QR-LMM is proven to be applicable for count data.

Recall the count mean mixed model or Poisson mixed model (Section (4.3)) for a discrete variable λ_{ij}, written as

$$\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + u_i), \quad i = 1, 2, ..., M; \quad j = 1, 2, ..., n_i,$$

where random effect u_i is assumed to be normally distributed with mean 0 and variance σ^2, $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$.

This mean model needs to be improved in order to estimate quantiles of y_{ij} given x_{ij} for a fixed $\tau \in (0,1)$, $Q_\tau(y_{ij}|x_{ij})$. This will be fulfilled by *jittering* the data. Machado & Silva (2005) had the idea of *jittering* in order to get continuous data as discussed in Subsection (4.2.4). *Jittering* the longitudinal count data can also be applied in the linear mixed model. The main idea is similar to the count data in linear models, see Machado & Silva (2005), for more details. By adding a standard uniform r.v. u_{ij} independent from y_{ij} and x_{ij}, a continuous observation z_{ij} can be obtained as

$$z_{ij} = y_{ij} + u_{ij}, \quad i = 1, 2, ..., M; \quad j = 1, 2, ..., n_i, \tag{4.36}$$

where y_{ij}'s are count observations, and $u_{ij} \sim$uniform$(0,1)$.

QR-LMM can be adapted on the continuous r.v. z_{ij} of Equation (4.32). Thus, the longitudinal quantile of the *jittered* data z_{ij} for a fixed $\tau \in (0,1)$ can be written as

$$Q_\tau(z_{ij}|x_{ij}) = \tau + \exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_\tau + \varepsilon_{\tau,ij}) \tag{4.37}$$

For a fixed $\tau \in (0,1)$ the quantile of the transformed *jittered* data $T(z_{ij},\tau)$ can be written as

$$Q_\tau(T(z_{ij},\tau)|x_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta}_\tau + \varepsilon_{\tau,ij}, \tag{4.38}$$

where

$$T(z_{ij},\tau) = \begin{cases} \log(\xi), & z_{ij} \leq \tau \\ \log(z_{ij}-\tau), & z_{ij} > \tau, \end{cases}$$

with a small value ξ. Thus, $T^{-1}(z_{ij},\tau) \approx \tau + \exp(z_{ij})$.

Since the transformed *jittered* data: $y^*_{ij} = T(z_{ij},\tau)$ is now continuous, as a result the quantile count estimation could apply in the linear mixed models. This will lead to the quantile estimator of y^*_{ij} given x_{ij} as

$$\hat{Q}_\tau(y^*_{ij}|x_{ij}) = \boldsymbol{x}'_{ij}\hat{\boldsymbol{\beta}}_\tau + \hat{\varepsilon}_{\tau,ij}, \quad i = 1,2,...,M; \quad j = 1,2,...,n_i \tag{4.39}$$

For a fixed $\tau \in (0,1)$ the estimator for the τ-quantile of the observed counts (back-transformed) y_{ij} given x_{ij} can be written as

$$\begin{aligned} \hat{Q}_\tau(y_{ij}|x_{ij}) &= \lceil T^{-1}(\hat{Q}_\tau(z_{ij}|x_{ij})) - 1 \rceil \\ &= \lceil \tau + \exp(\boldsymbol{x}'_{ij}\hat{\boldsymbol{\beta}}_\tau + \hat{\varepsilon}_{\tau,ij}) - 1 \rceil \end{aligned} \tag{4.40}$$

As in the QR model for independent data, the longitudinal quantile model works on continuous data. That is why the count data needed to be made continuous by Machado & Silva (2005) *jittering* method and a transformation in order to have a QR-LMM as in Equation (4.21). After the estimation, a back-transformation of the quantile estimators of the transformed *jittered* data will give the quantiles of the counts.

4.4 Data example: CAPRISA 002 AI Study data

In this section, the estimation of quantile mixed-effects models of Galarza et al. (2015) introduced in Section 4.3 is applied to the CAPRISA 002 Acute Infection Study data. The data set, which is a subset of the Centre for the AIDS Programme of Research in South Africa, consists of repeated CD4 count measurements and some other covariates of 235 individuals. Each subject has been measured several times, ranging from 2 to 61, with a median equal to 29. Table 4.1 illustrates a summary of the patients' baseline characteristics. The patients' age at enrollment ranges from 18-59, with the median age being 25 years. $Q_{0.05}$, which is a value that has 5% of the observation smaller or equal to it, indicates that 5% of the patients had a square root

of CD4 count below or equal to 16.4 at enrollment. $Q_{0.95}$ is also a value that shows 95% of the observation smaller or equal to it; said otherwise, 5% of the patients are greater than it. Therefore, Table 4.1 indicates 5% of the study participant had a square root CD4 count greater than 31.4 at enrollment. Moreover, the study participants had a mean BMI of 28.93 with minimum and maximum BMI of 17.89 and 54.89 at baseline. The median log baseline VL of the patients was 10.26 with minimum and maximum log baseline VL of 0 (Not detected) and 15.52, respectively (IQR = 2.91). Additional features on this dataset can be found in (Van Loggerenberg et al., 2008; Mlisana et al., 2014). We analyze this dataset intending to explain the different conditional distribution of the square-root-transformed CD4 count as a function of sets of covariates of interest by modeling a framework of response quantiles.

Table 4.1: Summary of patients' baseline characteristics

	Analysis						
Variables	**Mean**	**Median**	**Minimum**	**Maximum**	$Q_{0.05}$	$Q_{0.95}$	IQR
SQRT CD4 count	23.44	22.89	13.49	39.49	16.4	31.4	5.78
Baseline BMI	28.93	27.24	17.89	54.89	20	43.7	9.66
Log Baseline VL	10.09	10.26	0 (undetected)	15.52	6.19	13.13	2.91
Age at baseline	27.15	25	18	59	20	41	8

Based on the results of the information criteria, we compare four models. The comparisons of the models were made based on the 0.5^{th} quantile (median regression). The linear quantile mixed-effects model with random intercept and slopes (Model 4, see Table 4.2) was selected as the best model because the chosen model achieved the smallest Akaike information criteria (AIC), Bayesian information criteria (BIC), Hannan-Quinn information criteria (HQC), and the largest Log-likelihood (LL) (Table 4.2). Therefore, we examine the square-root-transformed CD4 count of HIV-infected patients as a response while accounting for baseline BMI, age, log baseline VL, and HAART initiation as predictor variables across various quantiles based on Model 4 (Table 4.3). To get a complete picture of the effects, a series of QR-LMM over the grid $\tau = \{0.25, 0.5, 0.75\}$ as well as estimation at $\tau = 0.05, 0.85$, and 0.95 are made.

Random effect models that were examined for analysis at 0.5^{th} quantile:
Model 1: Time
Model 2: Intercept, Time
Model 3: Time, $\sqrt{Time}$

Model 4: Intercept, Time, $\sqrt{Time}$

Table 4.2: Comparison of random effects models for QR-LMM at the 0.5th quantile

Random effects	AIC	BIC	HQC	LogLik
Model 1	39670.99	39725.84	39689.89	-19827.5
Model 2	35072.84	35141.41	35096.47	-17526.42
Model 3	35726.22	35794.79	35749.85	-17853.11
Model 4	**33685.92**	**33781.91**	**33718.99**	**-16828.96**

The linear mixed-effects model form of the data can be specified as:

$$y_{ij} = \beta_1 + \beta_2 t_i + \beta_3 \sqrt{t_i} + \beta_4 BMI_i + \beta_5 LVL_i + \beta_6 ART_i + \beta_7 Age_i + b_{1i} + b_{2i} t_i + b_{3i} \sqrt{t_i} + \varepsilon_{ij}$$

where y_{ij} is the transformed continuous form of CD4 count ($\sqrt{CD4count}$) at the j^{th} time point for the i^{th} subject, t is the time measured in months from the start of the study, BMI indicates the patient's baseline BMI, LVL= log of baseline VL, ART is the dichotomous HAART initiation (0 = pre-HAART, 1 = post-HAART), Age is patient's age at baseline, b_{1i} indicates the random intercept, b_{2i} and b_{3i} indicates the random slopes for subject i, and ε_{ij} the measurement error term, assumed ALD, for 235 subjects.

As can be viewed from Table 4.3, the intercept (β_1), which is the predicted value of the square-root-transformed CD4 count keeping all the other covariates constant, differ significantly across the quantiles, while time (β_2), square root of time (β_3), baseline BMI (β_4), the log of baseline VL (β_5), and post HAART initiation (β_6) significantly affect the CD4 count across all quantiles. In addition, although age (β_7) is found to have a positive and almost constant influence on the CD4 count across all quantiles, its effect is non-significant (Table 4.3). We can also see from Table 4.3; there is a remarkable positive effect of baseline BMI on the CD4 cell count ($\sqrt{CD4count}$) from low quantiles to higher quantiles. Whereas, from low to more upper quantiles, the negative effect of baseline VL on the count of CD4 cells increases gradually. This indicates that when the VL at enrollment is high (baseline VL at higher quantiles), its negative effect on the immune systems increases (Table 4.3). From low quantiles to upper quantiles, the post HAART initiation effect on CD4 cell counts has an increasing trend, and then at high quantile 0.95, its effect begins to decline.

Table 4.3: Parameter estimates for CAPRISA 002 AI study data across several quantiles

Parameter	$\hat{Q}_{0.05}(SE)$	$\hat{Q}_{0.25}(SE)$	$\hat{Q}_{0.5}(SE)$	$\hat{Q}_{0.75}(SE)$	$\hat{Q}_{0.85}(SE)$	$\hat{Q}_{0.95}(SE)$
β_1	19.99 (1.16)*	22.17 (1.4)*	24.63 (1.46)*	26.6 (1.42)*	27.97 (1.42)*	31.38 (1.39)*
β_2	0.06 (0.01)*	0.07 (0.01)*	0.06 (0.013)*	0.05 (0.013)*	0.04 (0.01)*	0.034 (0.015)*
β_3	-0.86 (0.14)*	-0.87 (0.13)*	-0.7 (0.11)*	-0.59 (0.12)*	-0.58 (0.124)*	-0.385 (0.16)*
β_4	0.05 (0.02)*	0.08 (0.02)*	0.082 (0.03)*	0.11 (0.03)*	0.13 (0.033)*	0.145 (0.031)*
β_5	-0.56 (0.08)*	-0.57 (0.1)*	-0.64 (0.09)*	-0.71 (0.09)*	-0.714 (0.08)*	-0.74 (0.084)*
β_6	1.68 (0.05)*	2.13 (0.07)*	2.56 (0.08)*	3.02 (0.09)*	3.114 (0.098)*	2.29 (0.089)*
β_7	0.021 (0.025)	0.03 (0.03)	0.03 (0.031)	0.029 (0.032)	0.026 (0.0321)	0.013 (0.03)
Log-lik	-18454.68	-17169.85	-16828.96	-17344.63	-17952.5	-19088.77
AIC	36937.36	34367.69	33685.92	34717.25	35933	38205.55

- Significance at 5% level. See, Additional outputs (7), for more significant test results and confidence intervals.

The results in graphical representation following QR-LMM over the framework of quantiles $\tau = \{0.05, 0.25, 0.5, 0.75, 0.85, 0.95\}$ have appeared in Figure 4.2. The graph shows that the 95% confidence interval for the covariates effect and the nuisance parameter σ. The figure reveals that the effect of baseline BMI (β_4) and post HAAR initiation (β_6) become more prominent across quantile levels, with their effect become more for higher conditional quantiles. In addition, although the effects of time (β_2) and baseline VL (β_5) exhibit a significant positive and negative influence, respectively, on the CD4 count across all quantiles, the difference changes with regard to the conditional quantile been more vital for lower quantiles. The $\hat{\sigma}$ is symmetric about $\tau = 0.5$, taking its maximum value at that point and decreasing for higher quantiles. The convergence of estimates for all parameters was also evaluated using the graphical criteria in Figure 4.3.

Some of the codes that were used for this section can be found here (Code 7.3 in the Appendix A).

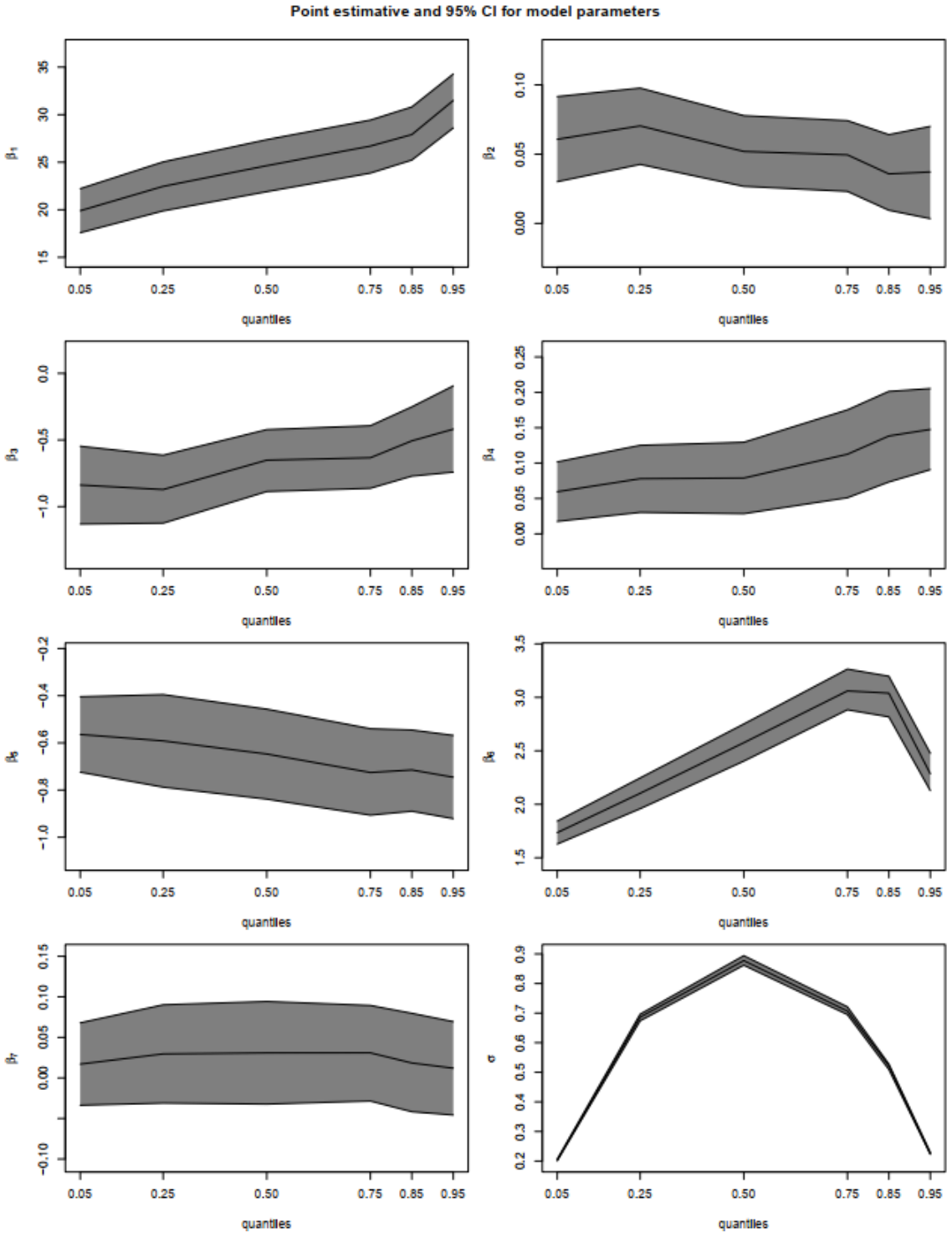


Figure 4.2: Point estimates and 95% confidence bands for model parameters following the QR-LMM to the CAPRISA 002 AI Study data across various quantiles

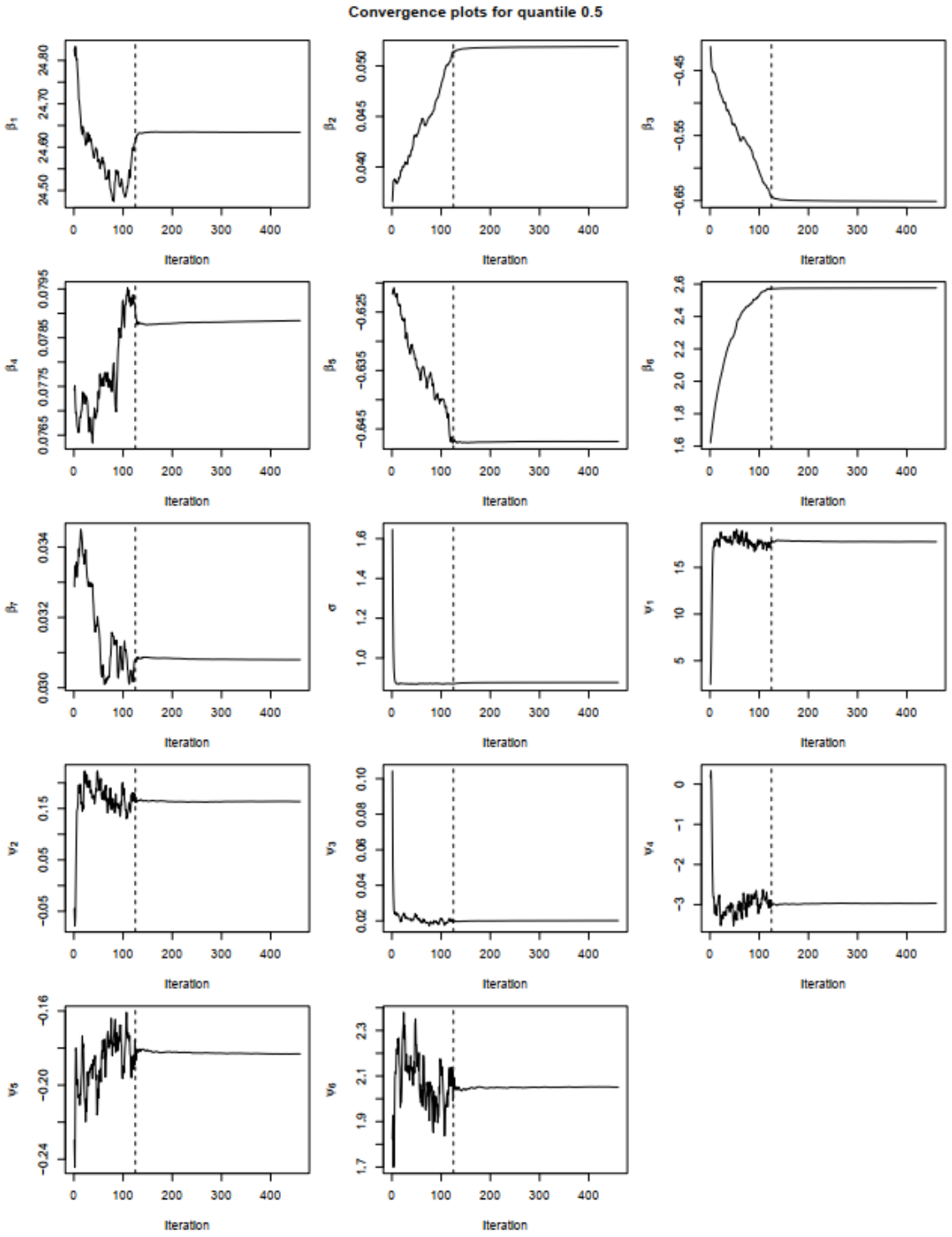


Figure 4.3: Graphic overview of convergence for model parameters at 0.5th quantile (as an example), produced from the qrLMM package using the CAPRISA 002 AI Study data

4.5 Summary

Mixed-effects models are well-established and popularly used approaches to examine the effect of longitudinal and/or repeated measurement data since they allow us to study within- and between-subject variation by taking into account the dependence of measurements of hierarchically structured data (Tian et al., 2020). However, mixed-effects models or more generally generalized mixed-effects models are mainly based on the distributional assumption of Gaussian, Poisson, negative binomial (Poisson-gamma mixture), and others for the outcome. Therefore, when the assumption is not attainable (sometimes even after the transformation of the outcome), the inferences based on these models are questionable. The ability to achieve promised normality may be possible at some point but could not be guaranteed over the full range of a relevant covariate such as age (Wei et al., 2006). Besides achieving the normality assumption for the outcome, resolving the extremely skewed data in applying mixed-effects models is another issue.

Further, mixed-effects models would be inadequate not only in estimating the location but also in estimating the percentile of the conditional distribution of the outcome while the sign of skewness of the distribution changes over the wide-ranging of the outcome (Geraci & Bottai, 2007). There are many ways to relax normality and overcome these issues. However, as is known, mixed-effects models focus mainly on the mean change of the outcome variable Y conditionally on the covariates X. Thus, applying them may result in non-robust estimation when the interest is in studying the effects across different quantile levels (outcome distribution) as well as for data with outliers and non-normal errors.

In contrast to mean-based regression models, the QR model, which belongs to a robust statistical model family, avoid the difficulty of these issues by giving an overall assessment of the covariate effects at different quantiles of the outcome and provides a complete picture of the relationships between the covariate and an outcome that are missed by other regression methods (Cade & Noon, 2003). QR methods dig deeper into the data, grab more information, and became more relevant.

QR apart from standard regression. By fitting models for more percentiles, one can detect the heterogeneous effects of covariates at the conditional distribution of the response, rather than just the conditional mean. That is especially useful when valuable information lies in the tails. QR estimates avoid distributional assumption forms such as those listed above for the random error term. However, the model's deterministic portion follows a parametric format (Cade & Noon, 2003). The con-

ventional mean-based regression estimates the model parameters by minimizing the sum of squares of residuals. In contrast, QR minimizes the sum of check loss functions of the residuals, and the estimates are depending on the quantile level τ. Thus, there is a distinct set of regression coefficients at each τ.

Recently, QR has also become practical for longitudinal and other forms of data due to the recent advances in computing resources and the ready availability of efficient LP algorithms, benefitting applications in various scientific areas (Huang et al., 2017). For independent count data, if all count models fail, one can use quantile count models that incorporate the idea of *jittering*. We illustrated the applicability of QR-LMM for longitudinal data. A series of QR-LMM over the grid $\tau = \{0.05, 0.25, 0.5, 0.75, 0.85, 0.95\}$ were estimated (Table 4.3), and the results were discussed.

Since quantile inference for discrete longitudinal data cannot thus be carried out directly yet, we followed the standard practice to model a continuous approximation of the quantile function by using square-root-transformed CD4 count as the response variable. Time since seroconversion, HAART initiation, and baseline characteristics of the patients such as BMI, age, and VL was included in the study. It was found that except age, all the studied variables significantly affected the count of CD4 cells of HIV-infected patients across all quantiles. Although significant CD4 cell recovery in response to post HAART initiation across all quantiles was recognized, HIV-infected patients who were enrolled in the treatment with a high level of VL showed a significant adverse effect on CD4 cell counts at upper quantiles.

Chapter 5

Analyzing longitudinal CD4 count of HIV-infected patients using generalized additive mixed-effects model

5.1 Introduction

Multiple linear regression models study linear relationships among two or multiple independent variables and one dependent (response) variable. We can extend the multiple linear regression model idea to the generalized linear model (GLM), where the distribution of the outcome variable can include distributions other than Gaussian. The response variable in GLM can be continuous, dichotomous, count, ordinal, categorical, and so on as long as its distribution is from an exponential family (Dobson & Barnett, 2008; McCullagh & Nelder, 1989). Consider the response variable whose domain is non-negative integer (count) values, which follows a Poisson distribution; if there is no over or under-dispersion, the mean and variance are assumed to be equal. However, the restriction (mean=variance) may not be satisfied with many real-data applications. Sometimes the variance is greater than the mean, and this phenomenon is called over-dispersion. One such model that works in such a condition is the negative binomial regression model (Hilbe, 2011, 2014; Yirga et al., 2020b). The negative binomial model is a generalization of the Poisson model, which relaxes the restrictive assumption that the mean and variance are equal. It has vast applications as a model for count data, especially for data showing over-dispersion (Hilbe, 2011, 2014; Yirga et al., 2020b).

The generalized linear model fails to consider the dependence of repeated observations over time. Therefore, it is necessary to extend the GLM to general linear mixed models. Linear mixed models (LMMs) characterize an ordinary and conventional type of regression method used to examine longitudinal studies data. The general form of LMM can be expressed as

$$y_{ij} = \beta_0 + \beta_{11}x_{111} + \cdots + \beta ipx_{ijp} + b_{i0} + b_{i1}z_{ij1} + \cdots + b_{ip}z_{ijp} + \epsilon_{ij} \tag{5.1}$$

where y_{ij} is an outcome variable that indicates the j^{th} measurement on the i^{th} subject, $x_{ijp}, j = 1, \ldots, n_i$ are the predictor variables, $\beta_0, \beta_1, \ldots, \beta_{ip}$ are fixed effects, $b_{i0}, b_{i1}, \ldots, b_{ip}$ are random effects, and ϵ_{ij}'s are random errors. If we want to generalize expression (5.1), we do not need to assume that the outcome variable is normally distributed. However, it has to follow a distribution from the exponential family. At that point, we can combine the idea of the mixed model with the GLM; hence, the resulted model is known as the generalized linear mixed model (GLMM) (Gbur et al., 2012; Stroup, 2012).

GLMMs include random effects into the linear predictor $g(\cdot)$ as an extension of GLMs. As an extension of the LMM, GLMMs contain fixed effects and random effects. This permits the modeling of correlated, conceivably non-normally distributed data. This may overcome the modeling issue of over-dispersion in the longitudinal and, at the same time, oblige the population heterogeneity (Gbur et al., 2012; Stroup, 2012). For these reasons, we used a negative binomial regression in the context of GLMMs to examine the CD4 count of HIV-infected patients as a function of HAART and other important factors parametrically in the previous study (Yirga et al., 2020b). More particularly, the GLMM has the following structure

$$g(E[y_{ij}|u_1, \cdots, u_q]) = \beta_0 + \sum_{i=1}^{p} \beta_{ij}x_{ij} + b_0 + \sum_{i=1}^{p} b_{ij}u_k, \tag{5.2}$$

where $y_{ij}, i = 1, \ldots, n; j = 1, \ldots, p$ is the outcome variable whose conditional distribution given the set of q random effects $(u_1, \ldots, u_q)$ belongs to the exponential family, x_{ij}'s are sets of p explanatory variables describing the fixed effects, and $g(\cdot)$ is the link function relating the conditional mean of the response to the predictors. The literature on GLM, LMM, and GLMM is ubiquitous, and one can find some of it here (Dobson & Barnett, 2008, 2018; McCullagh & Nelder, 1989; Pinheiro & Bates, 2006; Diggle et al., 2002; Demidenko, 2013; Gbur et al., 2012; Stroup, 2012; Wu & Zhang, 2006; Jones, 1993; Verbeke & Molenberghs, 2009; Diggle et al., 2002; Davidian & Giltinan, 2003; Diggle et al., 2002).

GLMMs incorporate nonlinear functional forms of the covariate effects as quadratic, square root, or cubic terms if these are thought to be necessary to provide an adequate fit (Der & Everitt, 2012; Lin & Zhang, 1999). This implicates that parametric regression models require the investigator to know in advance the functional form of the explanatory variables in the data. If the investigator knows that form, parametric regression models may be the appropriate choice (Shadish et al., 2014). However, the assumption of linear dependence of the outcome in parametric methods may not always be desired. In most cases, the associations between the outcome variable and explanatory variables may have an unidentified functional form. For such cases, the study of semiparametric additive mixed models becomes essential. Moreover, there will also be a complex form of relationships between the outcome variable and the covariates. As it nearly always is in real data analysis, the covariates' functional forms are rarely known (Melesse & Zewotir, 2020; Shadish et al., 2014). Also, parametric models suffer from inflexibility in many applications because they are too restrictive or limited; sometimes, it is challenging to find the proper parametric model (Wu & Zhang, 2006). Nonparametric regression methods have been developed to overcome these difficulties, where flexible, functional forms can be estimated from the data to capture possible complicated relationships between the response variable and multiple explanatory variables (Fitzmaurice et al., 2008; Wu & Zhang, 2006).

Nonparametric regression approaches allow the data to determine the model's appropriate functional forms, which best describe the available data (Fitzmaurice et al., 2008; Wu & Zhang, 2006; Ayele et al., 2014). This makes it important to fit a much larger class of models by reducing possible modeling biases of parametric models (Fitzmaurice et al., 2008; Wu & Zhang, 2006; Ayele et al., 2014). Nonparametric modeling relaxes the usual assumption of linearity and allows us to investigate the data more flexibly, revealing structures in the data that might otherwise be missed. However, when the number of covariates in the model is large, many forms of nonparametric approaches do not perform well. The inadequacy of data in this setting causes the variances of the estimates to be unacceptably large. The issue of rapidly increasing variance for increasing dimensionality is referred to as the "curse of dimensionality" (Xiang, 2001). Interpretability is also another issue with nonparametric techniques based on Kernel and Spline estimates, which are the most widely used estimator in nonparametric models. The information based on these estimates is often difficult to comprehend (Xiang, 2001). Stone (1985) proposed an additive model (AM) to overcome these difficulties. Thus, we first begin with the overview of AM then discuss the form of the negative binomial regression in the generalized additive mixed models setting in the next sections.

5.2 Additive models

The AM is a generalization of the nonparametric version of the multiple linear regression model. An AM with more than one explanatory variable can be expressed as

$$\boldsymbol{Y}_i = \boldsymbol{X}^*\boldsymbol{\beta} + \sum_{i=1}^{p} f_i(\boldsymbol{x}_i) + \varepsilon_i, \qquad \text{with} \qquad \varepsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2) \tag{5.3}$$

where $\boldsymbol{Y}_i$ is vector of response variable, $\boldsymbol{X}^*$ is a model matrix for all strictly parametric model components, β is the corresponding parameter vector, $f_i(\cdot)$'s are arbitrary univariate and smooth (nonparametric) functions one for each x_{ij}'s (covariates), and ε_i's are random errors (Hastie & Tibshirani, 1990; Hastie, 2017). In order to be estimable, the smooth functions x_i have to satisfy standard conditions such as $E(f_i(x_i)) = 0$. These functions are not given in a parametric form but instead are estimated in a nonparametric fashion (Xiang, 2001). Thus, the AM can deal with nonlinearity in covariates that are not the main interest in a study and 'adjust' for those effects appropriately (Der & Everitt, 2012).

Additive models assess an additive estimation of the multivariate regression methods. The advantages of an additive estimation are at least twofold. First, on account that each of the individual additive terms is assessed using a univariate smoother, the "curse of dimensionality" is prevented at the cost of not approximating universally. Second, the individual terms' estimates clarify how the dependent variable changes with the corresponding independent variables (Xiang, 2001).

5.2.1 Smoothing function

A smoother is a tool for summarizing a response measurement trend as a function of one or more predictor measurements $x_1, \cdots, x_p$. It provides an estimate of the trend that is less variable than the response variable itself. The vital property of a smoother is its nonparametric nature. It assumes a flexible form for the dependence of Y on $x_1, \cdots, x_p$. Hastie & Tibshirani (1990) provided a brief discussion of smoothers. With AMs, it is necessary to represent the smooth functions somehow and choose how they should be. Hastie & Tibshirani (1990) suggest representing AMs using spline-like penalized regression smoothers. Spline smoothing can be used to describe smooth functions so that expression (5.3) becomes a linear model. This is done by specifying a set of *basis functions* ϕ_{ij} for each function so that the smooth function can be represented as

$$f_i(x_i) = \sum_{j=1}^{q} \beta_{ij}\phi_{ij}(x_i) = \boldsymbol{\beta}_j'\boldsymbol{\phi}_j \tag{5.4}$$

where x_i's are covariates, basis functions ϕ_{ij} determine the spline, and β_{ij}'s are coefficients of the smooth, which will need to be estimated as part of the model set. Natural cubic spline, cubic smoothing splines, thin plate regression splines, and tensor product bases are some of the examples of penalized regression smoothers (Wood, 2017; Hastie & Tibshirani, 1990; Hastie, 2017).

5.2.2 Formulation and Estimation

There are many ways available to approach the formulation and estimation of AMs. The *backfitting* algorithm is a general algorithm that can fit an AM. The smooth functions $f_i(\cdot)$'s are fitted one at a time by taking the residual $y_j - \sum_{j \neq i} f_j(x_j)$. Then they are fitted against x_i using a smoother function. The process is repeated until it converges. Detailed discussion and formulation of the *backfitting* algorithm can be found here (Friedman & Stuetzle, 1981; Hastie et al., 2009; Chambers, 1991). Familiar tools for modeling and inference in multiple regression models are also available for AMs. While AMs are used in various statistical data analyses, there are types of issues for which they are not appropriate. For instance, the assumption of normality might not be adequate for modeling count outcomes, limitations for large data-mining applications, and the *backfitting* algorithm fits all predictors, which is not feasible or desirable when a large number are available (Xiang, 2001; Hastie & Tibshirani, 1990; Hastie et al., 2009; Hastie, 2017). Generalized additive models (GAMs) of Hastie & Tibshirani (1990) overcome these issues by extending AMs to a wide range of the exponential family of distributions, and not only the Gaussian. GAMs reduce to AMs when the outcome is normally distributed (Ruppert et al., 2003; Melesse & Zewotir, 2020) (Hastie & Tibshirani, 1990; Wood, 2017).

5.3 Generalized additive models

GAMs enable the response variable's mean to depend on an additive predictor through a nonlinear link function (Xiang, 2001). The GAMs combine an additivity assumption (Stone, 1985) that enables relatively many nonparametric relationships to be explored simultaneously and the distributional flexibility of GLMs (Nelder & Wedderburn, 1972). A GAM has the following general structure

$$g(\mu_i) = \boldsymbol{X}^* \boldsymbol{\beta} + \sum_{i=1}^{p} f_i(\boldsymbol{x}_i) \tag{5.5}$$

whereas usual x_i's are covariates, $\mu_i = E(Y_i)$ is the conditional mean of the response variable Y, which is linked to an additive function of the predictors through a link function $g(\cdot)$, and $f_i(\cdot)$'s are unspecified smooth (nonparametric) functions (e.g., cu-

bic smoothing spline, kernel smoothers, or thin-plate splines) (Hastie et al., 2009; Wood, 2003; Ayele et al., 2014). Note that the response variable Y is from the exponential family of distribution, and $g(\cdot)$ is a known monotonic twice differentiable link function (Hastie & Tibshirani, 1990). The GAMs are among those widely used nonparametric methods for independent data (Melesse & Zewotir, 2020; Ruppert et al., 2003; Chen, 2000). While the AM was estimated with penalized regression smoothers, the GAM is represented by penalized likelihood maximization, where the penalties are designed to suppress overly wiggly estimates of the f_i terms (Wood, 2017).

5.4 Additive mixed model

Longitudinal data such as repeated measures taken on each of several subjects frequently arises from many biological, ecological, and clinical studies and other scientific areas. Parametric mixed-effects models are powerful, well developed, parsimonious, and efficient tools, in particular, for modeling correlations and within/between-subject variations of longitudinal data when the models are correctly specified (Pinheiro & Bates, 2006; Demidenko, 2013; Diggle et al., 2002; Wu & Zhang, 2006; Jones, 1993; Verbeke & Molenberghs, 2009; Diggle et al., 2002; Davidian & Giltinan, 2003; Diggle et al., 2002). However, for many applications, as is said earlier, parametric models are often restrictive and less robust against model assumptions. For instance, in modeling the repeated outcome variable as a function of time and other covariates, the time effect is usually too complicated to be model parametrically. Thus, to relax these assumptions, nonparametric models have been developed for longitudinal data analysis, but they are usually more complex (Wu & Zhang, 2006; Müller, 2012). Semiparametric mixed-effects models (SMMs), which retain nice features of the mixed-effects modeling ideas and the nonparametric regression techniques, are a good compromise for longitudinal data analysis. A detailed discussion of SMMs can be found here (Ruppert et al., 2003; Wu & Zhang, 2006; Zeger & Diggle, 1994; Zhang et al., 1998; Tao et al., 1999; Durbán et al., 2005; Fan & Li, 2004; Harezlak et al., 2018).

Suppose that $y_{ij}(i = 1, \ldots, n; j = 1, \ldots, n_i)$ is the response for the i^{th} subject at time point t_{ij}, the SMM can be expressed as

$$y_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \sum_{i=1}^{p} f_i(\boldsymbol{x}_i) + \boldsymbol{z}'_{ij}\boldsymbol{b}_i + \sum_{i=1}^{p} \boldsymbol{U}_i(\boldsymbol{x}_i) + \varepsilon_{ij} \tag{5.6}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of coefficients associated with covariates $\boldsymbol{x}_{ij}$, $f_i(\cdot)$'s are twice-differentiable smooth functions of time or nonparametric fixed effects, $\boldsymbol{b}_i$ consists of independent $q \times 1$ vectors of random effects coefficients associated with covariates $\boldsymbol{z}_{ij}$, $\boldsymbol{U}_i(\cdot)$ is an independent and smooth random-effects' process and ε_{ij} is an independent measurement error at a time t_{ij} that is not clarified by either the fixed-effects component $(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \sum_{i=1}^{p} f_i(\boldsymbol{x}_i))$ or the random-effects component $(\boldsymbol{z}'_{ij}\boldsymbol{b}_i + \sum_{i=1}^{p} \boldsymbol{U}_i(\boldsymbol{x}_i))$ (Wu & Zhang, 2006; Zhang et al., 1998).

In general, SMM consists of the following four major components: parametric fixed-effects $(\boldsymbol{x}'_{ij}\boldsymbol{\beta})$, nonparametric fixed-effects $(f_i(\cdot))$, parametric random-effects $(\boldsymbol{z}'_{ij}\boldsymbol{b}_i)$, and nonparametric random-effects $(\boldsymbol{U}_i(\cdot))$. Wu & Zhang (2006) provided a detailed discussion of the different SMM types when one or two of the components expressed in the model (5.6) are dropped. For example, when only the nonparametric random-effects component is dropped, the SMM (5.6) reduces to expression (5.7), which is also the same as when the random-effects are incorporated into the AM (5.3), and it is referred to as additive mixed model (AMM)

$$y_{ij} = \boldsymbol{X}^*\boldsymbol{\beta} + \sum_{i=1}^{p} f_i(\boldsymbol{x}_i) + \boldsymbol{z}'_{ij}\boldsymbol{b}_i + \varepsilon_{ij} \tag{5.7}$$

where $\boldsymbol{X}^*$, $\boldsymbol{\beta}$, $(f_i(\cdot))$, $\boldsymbol{z}_{ij}$, $\boldsymbol{b}_i$, and ε_{ij} are defined as in (5.3) and (5.6); $\varepsilon_{ij} \sim N(0, \boldsymbol{R})$ and $\boldsymbol{b}_i \sim N(0, \boldsymbol{G}_\theta)$. Both covariate matrix $\boldsymbol{R}$ and $\boldsymbol{G}_\theta$ are positive-definite matrix depending on a parsimonious set of covariate parameters (Melesse & Zewotir, 2020; Ruppert et al., 2003; Zhang et al., 1998; Ayele et al., 2014). The AMM (5.7) is a hybrid extension of LMMs and AMs (Hastie & Tibshirani, 1990; Zuur et al., 2009; Mamouridis, 2011).

The AMM that is allowed to have some other distribution function and not only the Gaussian will be the generalized additive mixed models (GAMMs) (Melesse & Zewotir, 2020; Zuur et al., 2009; Mamouridis, 2011). A GAMM represents the model with higher flexibility and complexity, where LMM in which part of the linear predictor is specified as a sum of smooth functions of one or more predictor variables, and non-normally distributed outcomes are included (Melesse & Zewotir, 2020; Zuur et al., 2009; Mamouridis, 2011; Baayen et al., 2017; Lin & Zhang, 1999; Wood, 2017; Berhane & Tibshirani, 1998). Thus, a GAMM can be viewed as additive extensions of the GLMM (Melesse & Zewotir, 2020; Zuur et al., 2009; Mamouridis, 2011; Ayele et al., 2014; Wood, 2017).

5.5 Additive negative binomial mixed-effects model

Recall the negative binomial mixed-effects model that specifies the expected number of counts with mean μ_{ij} and parameter θ that controls over-dispersion discussed in our previous work Yirga et al. (2020b). We relate the conditional mean of the count response to the linear predictors through the logarithmic link function. As a result, we have

$$\begin{aligned} \log(\mu_{ij}) &= \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i \\ \mu_{ij} &= \exp\{\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i\} \end{aligned} \tag{5.8}$$

where $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ are known vectors of covariates associated with count data $y_{ij}, i = 1, \cdots, n$ and $j = 1, \cdots, n_i$, conditional on a q-dimensional vector of subject-specific random effects, $\boldsymbol{b}_i$, the counts of y_{ij}, with the assumption of gamma errors, has a negative binomial distribution, $y_{ij}|\boldsymbol{b}_i \sim NB(\mu_{ij}, \mu_{ij} + \theta\mu_{ij}^2)$ (Hilbe, 2011, 2014; Yirga et al., 2020b).

The additive negative binomial mixed-effects model replaces each or some of the linear term with a more general functional form

$$\begin{aligned} \log(\mu_{ij}) &= \boldsymbol{X}^*\boldsymbol{\beta} + \sum_{i=1}^{p} f_i(\boldsymbol{x}_i) + \boldsymbol{z}'_{ij}\boldsymbol{b}_i \\ \mu_{ij} &= \exp\{\boldsymbol{X}^*\boldsymbol{\beta} + \sum_{i=1}^{p} f_i(\boldsymbol{x}_i) + \boldsymbol{z}'_{ij}\boldsymbol{b}_i\} \end{aligned} \tag{5.9}$$

where again, each $f_i(\cdot)$ is an unspecified smooth function. While the nonparametric form for the functions $f_i(\cdot)$ makes the model more flexible, the additivity is retained and allows us to interpret the model in much the same way as GLMM form. The additive negative binomial mixed-effects model is an example of a GAMM (Zuur et al., 2009).

The general structure of GAMM can be written as follows

$$g(y_{ij}) = \boldsymbol{X}^*\boldsymbol{\beta} + \sum_{i=1}^{p} f_i(\boldsymbol{x}_i) + \boldsymbol{z}'_{ij}\boldsymbol{b}_i + \varepsilon_{ij}, \tag{5.10}$$

where y_{ij} is non-normally distributed outcome, $f_i(\cdot)$ is a centered twice-differentiable smooth function, $g(\cdot)$ is a monotonic differentiable link function, and $\boldsymbol{X}^*$, $\boldsymbol{\beta}$, $\boldsymbol{z}_{ij}$, $\boldsymbol{b}_i$, and ε_{ij} are defined as in (5.3) and (5.6). Statistical inference for GAMM involves inference of the nonparametric function $f_i(\cdot)$, which requires smoothing parameters and estimates of the variance components. In the Gaussian response and identity link function, the estimation of nonparametric functions, smoothing, and

variance parameters in GAMM are achieved using restricted maximum likelihood (REML) (Robinson et al., 1991; Silverman, 1985). For non-Gaussian response, penalized quasi-likelihood (PQL) (Breslow & Clayton, 1993) is the most widely used approach to estimate the parametric and nonparametric functions in GAMM (Lin & Zhang, 1999). A detailed discussion of PQL and several other approaches to estimate the smoothing parameters and variance components in GAMM can be found in numerous literature (Mamouridis, 2011; Kohn et al., 1991; Lin & Zhang, 1999; Zhang et al., 1998; Green & Silverman, 1993; Wahba, 1985; Wu & Zhang, 2006; Müller, 2012).

5.6 Data example: CAPRISA data set

5.6.1 Application of additive negative binomial mixed-effects model

In this section, the additive negative binomial mixed-effects model discussed in Section 5.5 is applied to the CAPRISA 002 AI Study data set. Tables 5.1 and 5.2 show the descriptive baseline measures of the dataset for this study. The dataset consists of a total number of 235 subjects. Each subject has been measured several times, ranging from two to sixty-one, with a total of 7129 observations. From a total of 235 women, 105 (44.7%) were residing around Vulindlela (rural area), and 130 (55.3%) were residing around eThekwini (Durban, urban area), KwaZulu-Natal, South Africa. The participants' age at enrollment ranges from 18-59 years, with the mean age being 27.15 years and a standard deviation of 6.56. The average CD4 count and viral load at enrollment were 570 (range 182-1575) with a standard deviation of 229.6 and 140442.31 (range 1 (undetected) - 5510000) with a standard deviation of 454895.893, respectively. Furthermore, the study participants had a mean BMI of 28.93 (range 17.89-54.89) with a standard deviation of 7.4 at enrollment. The majority of the women, 182 (77.4%), had a stable partnership, 224 (95.3%) completed secondary school, and most of them (78.8%) were self-reported sex workers (Mlisana et al., 2014; Van Loggerenberg et al., 2008; Yirga et al., 2020a,b).

Table 5.1: Baseline descriptive statistics for non-categorical variables

	Analysis			
Variables	**Mean**	**Std.Err**	**Minimum**	**Maximum**
CD4 cell counts (cells/μ L)	570	229.6	182	1575
HIV viral load (cells/mL)	140442.31	454895.893	1 (undetected)	5510000
Age (Years)	27.15	6.56	18	59
Body Mass Index	28.93	7.4	17.89	54.89

Table 5.2: Baseline descriptive statistics for categorical variables

Variable	Total	Variable	Total
Number of women	235	**Marital Status**	
Place of residence		No partner	43 (18.3%)
Rural	105 (44.7%)	Stable partner	182 (77.4%)
Urban	130 (55.3%)	Many partners	10 (4.3%)
Educational level			
Primary schools	11 (4.7%)		
Secondary schools	224 (95.3%)		

In our previous work, Yirga et al. (2020b), we fitted a parametric negative binomial mixed-effects model (NBMM) in the context of GLMM, assuming a linear relationship between the outcome and covariate. Now we aim to model the effect of time and some other covariates non-parametrically and incorporate parametric covariates using GAMM. Thus, in this study, we used an additive negative binomial mixed-effects model with a nonparametric time, age, and baseline BMI effect while the other covariates at hand remain parametric. The proposed model can be written as follows

$$\begin{aligned} g(\boldsymbol{Y}_i) = \beta_0 + \beta_1 \text{baseline VL}_i + \beta_2 \text{education}_i + \beta_3 \text{HAART}_i + \beta_4 \text{residence}_i \\ + \beta_5 \text{marital status}_i + f_1(\text{time in months}_i) + f_2(\text{age}_i) \\ + f_3(\text{baseline BMI}_i) + b_{0i} + b_{1i}(\text{time in months}_i) \end{aligned}$$

$$\boldsymbol{Y}_i \sim NB(\mu_{ij}, \mu_{ij} + \theta\mu_{ij}^2); \quad E(\boldsymbol{Y}_i) = \mu_{ij}; \quad Var(\boldsymbol{Y}_i) = \mu_{ij} + \theta\mu_{ij}^2$$

$$\begin{aligned} \mu_{ij} = \exp\{\beta_0 + \beta_1 \text{baseline VL}_i + \beta_2 \text{education}_i + \beta_3 \text{HAART}_i + \beta_4 \text{residence}_i \\ + \beta_5 \text{marital status}_i + f_1(\text{time in months}_i) + f_2(\text{age}_i) \\ + f_3(\text{baseline BMI}_i) + b_{0i} + b_{1i}(\text{time in months}_i)\} \end{aligned}$$

where $\boldsymbol{Y}_i$ is the vector of the response variable (number of CD4 cell), $g(\cdot)$ is the log link function, NB is a negative binomial distribution with mean= μ_{ij} and variance=$\mu_{ij}+\theta\mu_{ij}^2$, $\boldsymbol{\beta}_i$'s are parametric regression coefficients, $f_i(X)$ are a smooth function of the covariates X, and the random effects, $b_i \sim N(0, \boldsymbol{G}_\theta)$ (Ruppert et al., 2003; Lin & Zhang, 1999; Ayele et al., 2014; Melesse & Zewotir, 2020).

The R package *mgcv* was used to fit the above proposed model, using its *gamm* command (Wood & Wood, 2015). Significantly, the *gamm* command will penalize

'wiggly' lines to avoid overfitting, which suggests one can rappel all continuous covariates within smoothing functions. The model will then determine to what extent the data supports a 'wiggly' shape (Shadish et al., 2014). It also has several options available for controlling the model smoothness using splines. When the thin plate (tp) shrinkages splines were used to fit the above model, convergence was reached. Thin plate smoothers have the advantage of avoiding *knot* placement since they do not depend on the number of knots selected. They also provide computationally efficient and stable optimal approximations and can be constructed for smooths of more than one covariate at a time (Wood, 2003). Furthermore, the shrinkage smoothers obtained by using the bs option inside the s command are constructed so that smooth terms can be penalized away altogether, not contribute to the model (Wood, 2017; Zuur et al., 2009). The output is separated into parametric and smooth (nonparametric) parts. The smooth coefficients, $\boldsymbol{\beta}_i$'s are hidden inside the smoothers and are mostly uninterpretable. A smoother for the corresponding predictor can be fitted using the s function in the *gamm* command (Wood & Wood, 2015). The amount of smoothing of a smoother is expressed as effective degrees of freedom (edf), which essentially gives information on how 'wiggly' the fitted line is. A high value of edf ($\geq$8) indicates that the curve is highly nonlinear, whereas a smoother with edf $= 1$ means a linear relationship to the outcome (Zuur et al., 2009; Shadish et al., 2014).

Table 5.3 presents the log of expected CD4 count as parameter coefficients and the smooth terms' approximate significance using the proposed additive negative binomial mixed-effects model. Table 5.3 shows that baseline viral load and HAART initiation were found to have a significant effect on the progression of patients' CD4 count. The 'parametric coefficients' part showed that the patients' viral load at baseline had a significant adverse effect in the log of expected CD4 count, even if the change units are minimal. Moreover, the expected number of CD4 cells of the patient-initiated on HAART would be expected to multiply by 1.233 ($e^{0.2092}$) units compared to pre HAART initiation while holding other variables constant.

The results of edf from Table 5.3 shows that the variables age (edf=14.24, p-value $\leq$ 2e-16) and time (edf=10.343, p-value $\leq$ 2e-16) found to have a strong significant nonlinear effect on patients' CD4 count. The amount of smoothing for baseline BMI (edf=3.044, p-value=2.21e-06) indicates a significant nonlinear relationship with the response variable. The resulting fitted penalized spline plot is shown in Figure 5.1. The shaded region corresponds to pointwise approximate 95% confidence bands. The Y-axis indicates the smooth term's effect, where s is a smooth term, and the number in the parenthesis shows the edf value of the corresponding smooth term (Ayele et al., 2014). Visual inspection of Figure 5.1 shows that overall, the smoothers'

Table 5.3: Parameter estimates and approximate significance of smooth terms using an additive negative binomial mixed-effects model

Parameter coefficients	**Estimate**	**Std.Err**	**t-value**	**p-value**
Intercept	6.334e+00	1.172e-01	54.053	$\leq$ 2e-16
Baseline VL	-1.581e-07	4.709e-08	-3.358	0.00079
Educational level (ref.= Primary school)				
Secondary school	-1.500e-01	1.056e-01	-1.420	0.15564
HAART initiation (ref.= Pre HAART initiation)				
Post HAART initiation	2.092e-01	1.229e-02	17.021	$\leq$2e-16
Place of residence (ref.= Rural)				
Urban	3.569e-02	4.367e-02	0.817	0.41375
Number of sexual partner (ref.= No partner)				
Stable partner	4.490e-02	5.529e-02	0.812	0.41679
Many partner	-6.587e-02	1.116e-01	-0.590	0.55521
Approximate significance of smooth terms				
Smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(Age)	14.124	14.124	4.710	$\leq$ 2e-16
s(Time in months)	10.343	10.343	37.692	$\leq$ 2e-16
s(Baseline BMI)	3.044	3.044	9.759	2.21e-06

shape suggests that the progression of patients' CD4 count is higher after continuous follow-up time; the increment rate is low for the first four years (48 months) and steadily increasing afterward. The same relationships apply to the smooth terms age and baseline BMI; the CD4 count is higher for older patients and for those who have higher BMI at enrollment.

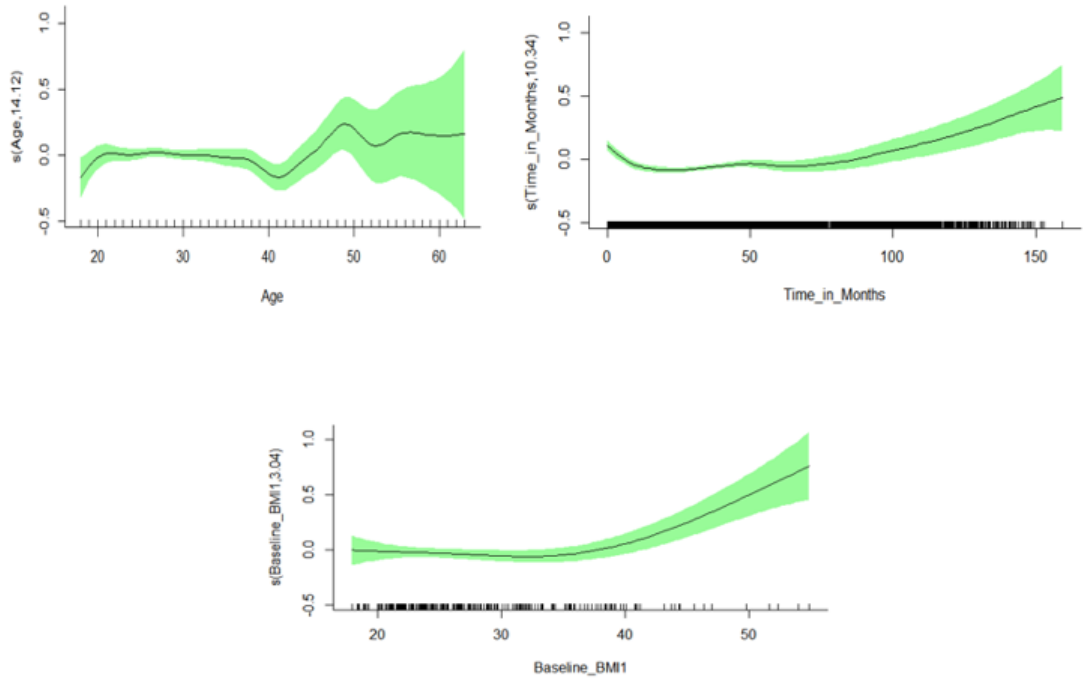


Figure 5.1: Estimated smooth curve for the GAMM model containing all smooth terms

To validate the fitted model, model diagnostics graphs were plotted (Figure 5.2). The normal Q-Q plot (upper left) shows some slight variation, but it is close to a straight line, suggesting that the distributional assumption is reasonable (Figure 5.2). The histogram of the residuals (lower left) is a bit more Gaussian. The plot of residuals versus fitted values (linear predictor), the upper right plot, shows that the variance is roughly constant as the mean increases. The response versus fitted values (lower left) plot suggests a positive relationship between the observed response and the fitted value (Figure 5.2).

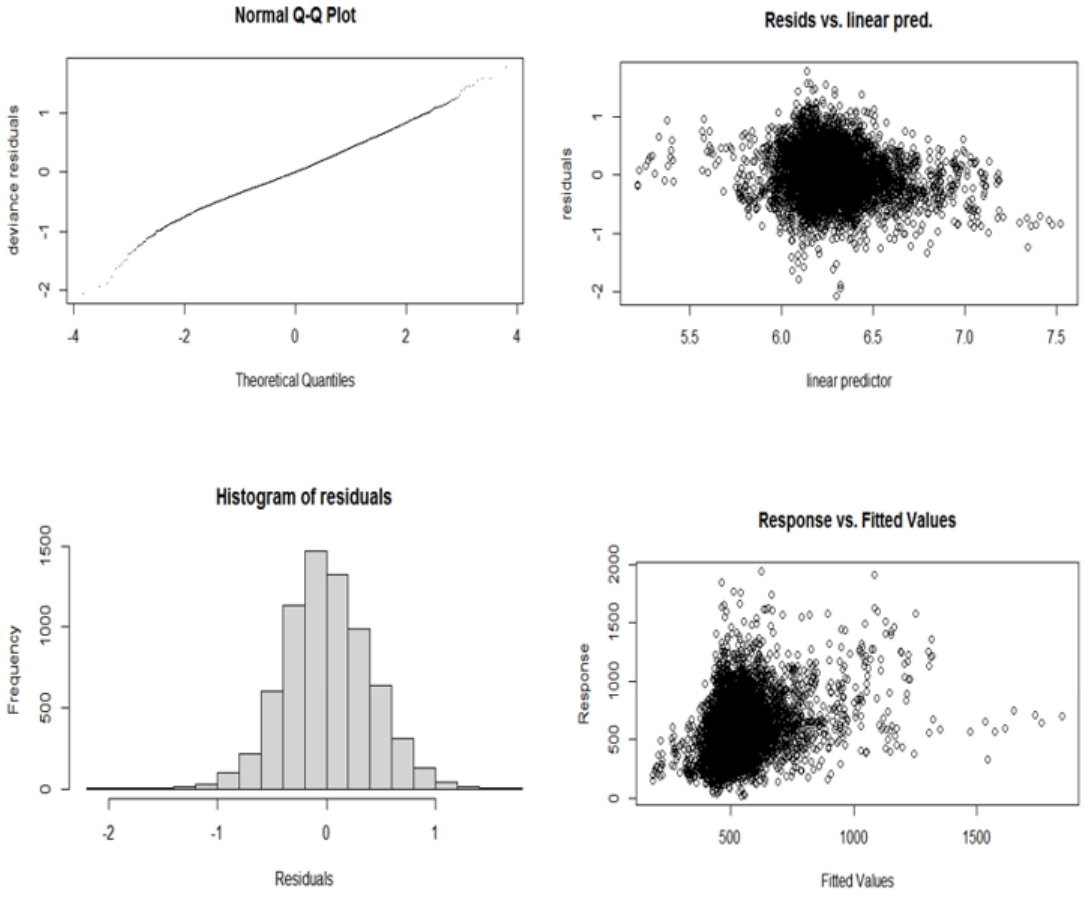


Figure 5.2: Diagnostic plots for checking the adequacy of the fitted model

Some of the codes that were used for this section can be found here (Code 7.4 in the Appendix A).

5.7 Summary

Multiple linear regression assumes that the relationship between the response (Y) and covariates (X) is linear or monotonic and constant across each variable's domain and range. However, there is no reason why every single regression must be linear or even have a particular structure, like being monotonic. To some extent, one can address this issue using polynomials. However, polynomials are not always desirable in terms of the model fit's properties; adding ever more powers of the covariate (X) to the model results in a model selection problem. Adding more powers to the covariate (X) in the polynomial model does not always improve the model's accuracy (Montgomery et al., 2021); instead, it could also result in a *Runge* phenomenon. Nonparametric regression approaches such as Locally Weighted Scatterplot Smoothing, sometimes called LOESS smoother, on the other hand, could be a better generalization since, in this method, there is no restriction on the functional form between the outcome and the covariates, except that the functional form has to be smooth. This means, if there is no restriction, the fits will be more computationally intensive. However, if LOESS smoothers are correctly done, they give us extra information from the data, but the information we get depends on the smoothing parameter's correct choice; the same applies to kernel smoothing. GAMs do offer a solution to these issues. They give us a framework to model flexible nonlinear relationships in the data.

GAM is a generalization of the multiple linear regression model as also of the GLM, where one can continue to model outcomes that arise from the exponential family such as continuous, discrete, counts, proportions, and so on. GAMs are versatile models used to understand better and analyze complex, nonlinear relationships in the data. They can capture critical aspects of the relationship between the response variable and the covariates by fitting the data with smooth or splines, which are functions that can take on a wide variety of shapes. One can fit GAM using the *gam* function from the *mgcv* package in R. When fitting GAM, the covariate (X) has to rappel in the s (smooth) function to specify the relationship to be flexible. A GAM can capture various nonlinear aspects because of the flexibility of these splines. The flexible smooths in GAMs are constructed of many smaller functions; these are called *basis functions*. Each smooth is the sum of several basis functions, and each basis function is multiplied by a coefficient, each of which is a parameter in the model. One can use GAMs to perform a multiple regression model that contains a mixture of smooth, linear effects, continues, and counts, or categorical variables. Not every term in a GAM has to be nonlinear. GAMs allow us to combine linear and nonlinear terms; to add a linear term, we do not have to rappel the predictor term in the s

function. Linear terms are useful when we have categorical variables as predictors in GAM.

GAMM, a mixed-effects version of GAM, is the most powerful model to deal with nonlinear trajectories in the longitudinal data. In this study, we have used an additive negative binomial mixed-effects model, an example of a GAMM, to analyze the longitudinal CD4 count of HIV-infected patients as a function of time, age, and baseline BMI non-parametrically as well as some covariates at hand parametrically. The analysis shows that the progression of CD4 count of the patient is significantly increased after the patient had been initiated on HAART, and the baseline viral load of the patient has shown a significant adverse effect in the progression of their CD4 count over time, as we would have expected. Our analysis also identified a significant nonlinear effect of age, baseline BMI, and time. The results from the nonparametric part of the model revealed that the progression of CD4 count is high for older aged ($\geq$ 40 years old) patients. Moreover, patients with higher BMI at baseline have shown improvement from the treatment over time. However, it does not mean that patients with higher BMI should be clinically ignored. Instead, the study confirms that BMI contributes to drug metabolism and consequently influencing the progression and immunological responses of HAART. Furthermore, the significant nonlinear time effect has shown patients' CD4 count progression is low; the progression is getting starting after many treatment visit times. Thus, the study suggests that effective HAART initiation after HIV exposure is necessary to suppress the increase of viral loads to induce potential ART benefits that accrue over time, especially during the COVID-19 infection since evidence are showing that HIV patients who are not clinically and immunologically stable on HAART may be at higher risk of developing illness if they are infected with the coronavirus.

Chapter 6

Additive quantile mixed effects modelling with application to longitudinal CD4 count data

6.1 Introduction

In this chapter, the topics nonparametric and additive quantile regression models are discussed. We also introduced and applied the additive quantile mixed model on real data sets as a general method for longitudinal data that has recently gained popularity. Parametric models relate the mean of a response variable to a linear combination of covariate effects and focus on the response's average properties (Fenske et al., 2011). Nevertheless, there are inevitable occasions when such parametric models fail, and data analysis must turn to more flexible, nonparametric models (Koenker, 2005a). Parametric models also assume a distribution for the outcome variable as opposed to purely nonparametric models. However, most of the vast literature on nonparametric regression also deals with the estimation of conditional mean models. In addition, the conventional assumption of nonparametric regression theory that there is additive, independently, and identically distributed (*iid*) error around a smooth underlying conditional mean function is highly implausible in certain data settings (Koenker, 2005a). Thus, as in the parametric context, nonparametric methods are usefully complemented by nonlinear estimation of families of conditional quantile functions that relax the independence assumption (Koenker, 2005a). The use of parametric and nonparametric regression models for analyzing patients' CD4 count in most applications implies that the estimated effects describe the average CD4 count. However, it is of even great interest to examine the quantile

of the outcome distribution, such as the lower ($\leq 25\%$) quantile, which identifies patients at higher risk of developing illnesses.

As discussed in Chapter 4, quantiles, commonly symbolized by the Greek letter τ, are location and scale parameters simultaneously. For a given $\tau \in (0, 1)$, the τ^{th} quantile is the value of a random variable, where $\tau \times 100\%$ of its value lies below it. In other words, it is the value where at most $(1 - \tau) \times 100\%$ of the value lies above. Thus, τ^{th} quantiles close to 0.5-quantile give the median, which is a well-known location parameter. On the other hand, τ^{th} quantiles close to zero or one give an idea of the scale. For instance, the interquartile range (IQR) is defined as the first quartile subtracted from the third quartile: $IQR = Q_3 - Q_1$.

Quantile regression (QR) solutions are computed for a selected number of quantiles, typically the three quantiles along with two extreme quantiles, that is, for $\tau = \{0.05, 0.25(Q_1), 0.5(Q_2), 0.75(Q_3), 0.95\}$. This necessitates the search for a suitable compromise between the amount of output to manage and the results to interpret and summarize (Davino et al., 2013). Although in many practical applications of QR, the focus is on estimating a subset of quantiles, however, it is worth noticing that it is possible to obtain estimates across the entire interval of conditional quantiles; in particular, the set: $\{\beta_\tau : \tau \in (0, 1)\}$ (Koenker, 2005a).

QR is a versatile statistical method with many applications that complement mean regression (Koenker & Bassett, 1978; Geraci, 2019). Thus, it emerged as an effective analytic technique in numerous study areas of science due to its competence to drive inferences about individuals that rank below or above the conditional population mean and/or focused on features of the response beyond its central tendency (Buchinsky, 1998; Peterson & Krishnan, 2015; Sherwood et al., 2013; Yu et al., 2003; Cade & Noon, 2003; Yirga et al., 2018; Koenker & Geling, 2001; Koenker et al., 2011; Fenske et al., 2011; Geraci, 2019). QR is specifically appropriate for the parameters' heterogeneous effect as it yields inferences that can be legitimate irrespective of the true underlying distribution (Winkelmann, 2006; Geraci, 2019). QR techniques look further into the data, get more information, and become more important (Huang et al., 2017). By fitting models for more percentiles, one can detect the covariates' heterogeneous effects at the conditional distribution of the response, rather than just the conditional mean. That is especially useful when valuable information lies at the bottom or top quantiles. "QR also enjoys several properties, including equivariance to monotone transformations and robustness to outliers" (Koenker, 2005b; Gilchrist, 2000). A semiparametric extension of quantile regression models with different types of nonlinear effects included in the model equation leads to an additive

quantile regression model (AQM) (Fenske et al., 2011). Such a model may reveal systematic differences in dispersion, tail behavior, and other features with respect to covariates (Koenker, 2005a).

6.2 Nonparametric quantile regression

In a parametric regression model, the function connecting the response variable's conditional values to the covariates is a prior known and fixed as a linear function. However, in real data applications, such a linearity assumption might be substantial and lead to one-sided results. Nonlinear assumptions between study variables occur in many research studies (Wu & Zhang, 2006; Fitzmaurice et al., 2008; Lindsey, 2001; Davidian & Giltinan, 2003). In the process of nonlinearity, there are various modeling techniques one may consider. Nonparametric models, smoothing splines, and transformation models are the most adopted strategies that consider analytical framework such as types of sampling design (cross-sectional or longitudinal), kinds of outcome (discrete or continuous), distributional assumptions (parametric or nonparametric), and so on (Geraci, 2019). The effort required for the investigation might have significant weight on the ultimate choice concerning which method to follow. Lack of theory or computer programming can also move the needle towards one decision over another (Geraci, 2019).

Nonparametric regression permits the presumption of linearity to be relaxed (Wu & Zhang, 2006; Fitzmaurice et al., 2008; Fox, 2000) and limits the analysis to smooth and continuous functions (Davino et al., 2013). Nonparametric regression, also known as scatter smoothing, aims to distinguish the best regression function according to the data distribution instead of estimating the parameters (Davino et al., 2013).

Recall the nonparametric regression model:

$$y = \sum_{i=1}^{n} f_i(x_i) + \varepsilon_i,$$

where the function $f_i(\cdot)$ is unknown, and commonly assumed that the errors are normally and identically distributed: $\varepsilon_i \sim NID(0, \sigma^2)$ (Davino et al., 2013). Several methods have been introduced to model nonparametric regression models; however, the most used techniques that have been extended to QR are local polynomial regression (Chaudhuri, 1991) and smoothing splines (Hastie & Tibshirani, 1990; Hendricks & Koenker, 1992): for further details, see Wu & Zhang (2006), Davino et al. (2013), Fox (2000), Craig & Ng (2001), Koenker et al. (1992), Koenker et al.

(2008), Cleveland & Loader (1996), and Koenker et al. (1994).

Recall the parametric QR model, which is given by

$$Y_i = \boldsymbol{x}_i'\boldsymbol{\beta}_{\tau i}, \qquad i = 1, ..., n, \qquad 0 < \tau < 1,$$

where Y_i is the response variable, x_i's are covariates, $\beta_{\tau i}$'s are the quantile specific linear effects, and $\varepsilon_{\tau i}$ is a random variable assumed to be an unknown error term on which no specific distributional assumptions are made except that the distribution is restricted to have the τ^{th} quantile to be zero (Liu & Bottai, 2009; Lachos et al., 2015; Fenske et al., 2011). For this reason, the parametric QR model aims at describing the quantile function $Q_{Y_i}(\tau|\boldsymbol{x}_i)$ of the continuous outcome Y_i conditional on covariate vector x_i at a given quantile τ, and this can be expressed as follows

$$Q_{Y_i}(\tau|\boldsymbol{x}_i) = F_{Y_i}^{-1}(\tau|\boldsymbol{x}_i) = \boldsymbol{x}_i'\boldsymbol{\beta}_{\tau i} + \varepsilon_{\tau i}, \qquad \text{with} \qquad Q_{\varepsilon_{\tau i}}(\tau|\boldsymbol{x}_i),$$

where $F_{\tau i}$ is subject to $F_{\tau i}(0) = \tau$, $F_{Y_i}^{-1}(\cdot)$ is the inverse cumulative distribution function of Y_i. For a comprehensive overview of QR, see, for example, Koenker (2005b), Koenker & Hallock (2001), Koenker & Bassett (1978), Buchinsky (1998), or Yu et al. (2003).

As much as the parametric QR assumptions enjoy a simple model structure, convenience of interpretation, and lower computational cost, it is not flexible enough and hence carries the risk of model misidentifications for complex problems (Lin et al., 2013). Nonparametric QR has become a viable alternative to avoid restrictive parametric assumptions. Koenker et al. (1994) explored nonparametric QR in spline models (quantile smoothing splines), which they defined as solutions to

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \rho_\tau(y_i - f(x_i)) + \lambda \left(\int_0^1 |f''(x)|^p dx \right)^{1/p}, \tag{6.1}$$

where $\rho_\tau(u) = u\{\tau - I(u < 0)\}$, $p \geq 1$, is the so-called check (loss) function, the parameter $\tau \in (0, 1)$ controls the quantile of interest, and $\lambda \in R^+$ is a smoothing parameter (Koenker & Bassett, 1978; Koenker et al., 1994).

As closely analogous to the parametric QR model (4.7), Koenker (2005a) generalized the nonparametric QR models as

$$Q_{Y_i}(\tau|\boldsymbol{x}_i) = f(\boldsymbol{x}_i, \beta_i(\tau)), \tag{6.2}$$

Then, Koenker (2005a) presented the τ^{th} nonparametric QR estimator as

$$\hat{\boldsymbol{\beta}}_i(\tau) = \arg\min_{\beta} \sum_{i=1}^{n} \rho_\tau(y_i - f(\boldsymbol{x}_i, \beta_i(\tau))) \tag{6.3}$$

Several techniques were proposed for nonparametric QR modelings, such as Bivariate quantile smoothing spline (He et al., 1998) and Kernel quantile regression (Li et al., 2007). However, nonparametric QR is an important yet challenging topic that needs to be addressed in-depth (Lin et al., 2013). One can find a brief account of nonparametric QR strategies in numerous studies; see, for example, Davino et al. (2013), and Koenker (2005a). To account nonlinearity relationship between quantiles of the outcome and covariates, Rigby & Stasinopoulos (2005) also proposed generalized additive models for location, scale, and shape (GAMLSS). GAMLSS enables additional flexibility to fit the covariates' nonlinear effect; however, they do not result in easily interpretable expressions for the quantiles. They are based on specifying distinct distributional parameters (Fenske et al., 2011). Instead, additive quantile regression models (AQMs) allow for the inclusion of nonlinear covariate effects and give more flexibility (Fenske et al., 2011).

6.3 Additive quantile regression

As we also discussed it in Chapter 5, additive models, introduced by Stone (1985), Breiman & Friedman (1985), and Hastie & Tibshirani (1990), are flexible regression tools that manipulate linear as well as nonlinear terms. The nonlinear terms in additive models are modeled through smoothing splines (Geraci, 2019). They provide programmatic approaches for nonparametric (nonlinear in parameters) regression modelings; by restricting nonlinear covariate effects to be composed of low-dimensional additive pieces so that we can overcome some of the worst aspects of the notorious curse of dimensionality (Koenker et al., 2011). The literature on additive models is vast (Hastie & Tibshirani, 1990; Stone, 1985; Der & Everitt, 2012; Xiang, 2001; Wood, 2017). However, most of the work has been done based on estimating conditional mean functions. The additive quantile regression model (AQM) provides an attractive framework for parametric as well as nonparametric regression illustrations focused on features of the response beyond its central tendency (Koenker et al., 2011; Fenske et al., 2011; Geraci, 2019).

Fenske et al. (2011) defined the τ^{th} AQMs that extend the linear predictor, $\boldsymbol{x}_i'\boldsymbol{\beta}_\tau$, with a sum of nonlinear functions of continuous covariates, $\sum f_{\tau j}(\cdot)$, as follows

$$Q_{Y_i}(\tau|\boldsymbol{x}_i, \boldsymbol{z}_i) = \boldsymbol{x}_i'\boldsymbol{\beta}_{\tau i} + \sum_{j=1}^{q} f_{\tau j}(\boldsymbol{z}_i) + \varepsilon_{\tau i}, \qquad j = 1, ..., q, \tag{6.4}$$

where $f_{\tau j}$ denote generic functions of covariates z_i for the i^{th} observation, and allows for the inclusion of different model terms such as nonlinear effects (smooth functions) of z_k, $f_\tau(z_k)$, and varying coefficient terms, $z_k' f_\tau(z_k)$, where the effect of the covariate z_k' varies smoothly over the domain of z_k according to some function of f_τ. However, the underlying assumption of the error term, $\varepsilon_{\tau i}$, remains the same as in the QR model (4.7); see Fenske et al. (2011) for more details.

AQM estimates the additive effect using linear programming algorithms as in the conventional QR model (Fenske et al., 2011). However, in the AQM case, determining adequate numbers and the position of knots is challenging. To avoid these challenges, Fenske et al. (2011) used penalty methods such as quantile smoothing splines of Koenker et al. (1994). Thus, the minimization problem of AQM that consists of extra penalty term is given by (Fenske et al., 2011):

$$\arg\min_{f_\tau} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_{\tau i} - \sum_{j=1}^{q} f_{\tau j}(\boldsymbol{z}_i)) - \lambda \mathcal{V}(f_\tau'), \tag{6.5}$$

where $\mathcal{V}(f_\tau') = \sup \sum_{i=1}^{n-1} |f_\tau'(z_{i+1}) - f_\tau'(z_i)|$, which represents the total variation of the derivation $f_\tau' : [a, b] \to \mathbb{R}$, where the *sup* is taken over all partitions $a \leq z_1 < \cdots < z_n < b$, and λ is a tuning parameter that controls the smoothness of the estimated function also known as "total variation regularization": see Fenske et al. (2011), Koenker et al. (1994), Koenker et al. (2011) or Koenker (2005b) for more details.

6.4 Additive quantile mixed model

Additive mixed models (AMMs), an extension of additive models, have been developed precisely to incorporate linear and nonlinear effects, as well as random terms when the data are sampled according to longitudinal designs (Wood, 2017; Geraci, 2019). AMMs have been integrated into QR methods to obtain robust results, not only focused on features of the longitudinal outcome at its central tendency that may not be the best location to characterize the data specifically when the errors are non-normally distributed, and the location-shift hypothesis of the normal model is

violated but also at conditional quantiles of the longitudinal outcome with no assumption about the response or errors distribution apart from the distribution is restricted to have the τ^{th} quantile to be zero (Fenske et al., 2011; Liu & Bottai, 2009; Lachos et al., 2015). Thus, additive quantile mixed models, which have gained popularity recently as a general method for longitudinal data, bring a comprehensive and more complete picture of the nonparametric as well as the parametric effects (Fenske et al., 2013; Geraci, 2019).

Fenske et al. (2013) proposed extending AMMs to the QR model for longitudinal data that consists of fixed individual-specific intercepts and slopes modeled through penalized splines of Ruppert et al. (2003). However, their model did not include random-effects terms and did not allow individual-specific effects to have a general covariance structure (Geraci, 2019). The version of Geraci (2019) additive QR model for longitudinal data includes linear and nonlinear terms, as well as multiple random effects to account for the correlation at the individual level with a general variance-covariance matrix and allow for automatic smoothing selection within a mixed model framework of Ruppert et al. (2003). Thus, as pointed out by Geraci (2019), because of the following two basic ideas, his model shown to have superior performance compared with the approach of Fenske et al. (2013): the first point is regarding the i^{th} unit effects, which he assumed to be random instead of fixed so that the covariance structure between effects can be introduced; the second point is that instead of prior specification, the nonparametric term's smoothing is automatically estimated from the data (Geraci, 2019).

Geraci (2019) defined the τ^{th} additive QR model for longitudinal data as

$$\begin{aligned} Q_{y_{ij}|\boldsymbol{u}_i,\boldsymbol{x}_i,\boldsymbol{z}_i}(\tau) = \beta_{\tau,0} + \sum_{k=1}^{p} f_\tau^k(x_{ijk}) + z'_{ij}u_{\tau,i}, \\ j = 1, ..., n_i, \quad i = 1, ..., m, \quad \tau \in (0,1), \end{aligned} \tag{6.6}$$

where x'_{ij} is the j^{th} row of a known $n_i \times p$ matrix $\boldsymbol{X}_i$, z'_{ij} is the j^{th} row of a known $n_i \times q$ matrix $\boldsymbol{Z}_i$, y_{ij} is the j^{th} observation of the response vector $\boldsymbol{y}_i = (y_{11}, ..., y_{1n_i})'$ for the i^{th} unit, $f_\tau^k(\cdot)$ is a τ-specific, centered, twice-differentiable smooth function of the k^{th} component of $\boldsymbol{x}$, and $u_{\tau,i}$ is a $q \times 1$ vector of values that collects i^{th} unit random effects associated with z_{ij} and its distribution is assumed to depend on a τ-specific parameter (Geraci, 2019).

Geraci (2019) considered a spline model of the type: $f_\tau(x) \approx \sum_{h=1}^{H} \mathcal{V}_{\tau,h} B_h(x)$, to model nonlinear functions of the components of $\boldsymbol{x} = (x_1, ..., x_s, x_{s+1}, ..., x_p)'$ that

consists of the first s terms of nonlinear functions and $p - s$ linear functions. The B_h's denote the *basis functions* ($\mathcal{V}_\tau$), h's represent the corresponding τ-specific coefficients of B_h's and H indicates the number of knots (Geraci, 2019). The approximated quantile function from the model (6.6) is then expressed as follows (Geraci, 2019):

$$Q^*_{y_{ij}|\boldsymbol{u}_i,\boldsymbol{x}_i,\boldsymbol{z}_i}(\tau) = \beta_{\tau,0} + \sum_{k=1}^{s}\sum_{h=1}^{H_k} \mathcal{V}_{\tau,hk}\, B_h^{(k)}(x_{ijk}) + \sum_{k=s+1}^{p} \beta_{\tau,k}\, x_{ijk} + z'_{ij} u_{\tau,i} \tag{6.7}$$

In matrix notation, the i^{th} unit of expression (6.6), which is then called additive quantile mixed model (AQMM), is given by (Geraci, 2019)

$$Q^*_{y_{ij}|\boldsymbol{u}_i,\boldsymbol{x}_i,\boldsymbol{z}_i}(\tau) = \boldsymbol{F}_i\boldsymbol{\beta}_\tau + \boldsymbol{Z}_i\boldsymbol{u}_{\tau,i} + \boldsymbol{B}_i\boldsymbol{\mathcal{V}}_{\boldsymbol{\tau}}, \tag{6.8}$$

where $B^{(k)}(x_{ijk})$ is considered as $H_{k\times 1}$ vector of values taken by the k^{th} spline evaluated at x_{ijk}, $\mathcal{V}_{\tau,k} = (\mathcal{V}_{\tau,1}, ..., \mathcal{V}_{\tau,H_k})'$ considered as the $H_{k\times 1}$ vector of spline coefficients for the k^{th} covariate, and $H = \sum_k H_k$. Furthermore, $\boldsymbol{B}_i$ and $\boldsymbol{\mathcal{V}}_{\boldsymbol{\tau}}$, defined, respectively, as the $n_i \times H$ matrix with rows $\left(B^{(1)}(x_{ij1})', ..., B^{(s)}(x_{ijs})'\right)'$ and $(\mathcal{V}'_{\tau,1}, ..., \mathcal{V}'_{\tau,s})'$, $\boldsymbol{F}_i$ is the $n_i \times (p - s + 1)$ matrix with rows $(1, x_{ij(s+1)}, ..., x_{ijp})'$ and $\boldsymbol{\beta}_\tau = (\beta_{\tau,0}, \beta_{\tau,s+1}, ..., \beta_{\tau,p})'$ (Geraci, 2019).

The objective function of AQMM, where the vectors $\boldsymbol{u}_{\boldsymbol{\tau},\boldsymbol{i}}$ and $\boldsymbol{\mathcal{V}}_\tau$ assumed to follow zero-centered multivariate Gaussians with variance-covariance matrices $\sum_\tau$ and $\boldsymbol{\Phi}_\tau = \bigoplus_{k=1}^{s} \phi_{\tau,k}\, I_{H_k}$, respectively, with selecting $\rho_\tau(\boldsymbol{r}) = \sum_{j=1}^{n} r_j\{\tau - I(r_j < 0)\}$ for a vector $\boldsymbol{r} = (r_1, ..., r_n)'$, is given by Geraci (2019) as

$$\sum_{i=1}^{M} \rho_\tau(\boldsymbol{y}_i - \boldsymbol{F}_i\boldsymbol{\beta}_\tau - \boldsymbol{Z}_i\boldsymbol{u}_{\tau,i} - \boldsymbol{B}_i\boldsymbol{\mathcal{V}}_{\boldsymbol{\tau}}) + \sum_{i=1}^{M} \|u_{\tau,i}\|^2_{\sum_\tau^{-1}} + \sum_{k=1}^{s} \phi_{\tau,k}^{-1} \|\mathcal{V}_{\tau,k}\|^2, \tag{6.9}$$

where "$\boldsymbol{u}_{\tau,i}$'s are assumed to be independent for different i (but may have a general covariance matrix) and are independent from $\boldsymbol{\mathcal{V}}_\tau$, and $\phi_{\tau,k}$'s are determine the amount of smoothing for the nonparametric terms" (Geraci, 2019). Minimizing the objective function of expression (6.9) proceeds as the same as minimizing the objective function of quantile mixed-effects models (Geraci & Bottai, 2007; Galarza et al., 2015; Lachos et al., 2015) where the asymmetric Laplace distribution with a location parameter μ, scale parameter $\sigma > 0$, and skewness parameter $\tau \in (0, 1)$ (Koenker & Machado, 1999; Yu & Moyeed, 2001; Geraci & Bottai, 2007; Yu & Zhang, 2005), employed as *quasi-likelihood* for fidelity term (Geraci, 2019). Further discussion of AQMM is provided by Geraci (2019).

6.5 Data example: Subset of the CAPRISA study data

In this section, we illustrate the use of the AQMM of Geraci (2019) introduced in Section 6.4 on the Centre for the AIDS Programme of Research in South Africa (CAPRISA) 002 Acute Infection Study data. As we mentioned in the previous chapters, the CAPRISA study was effected at the Doris Duke Medical Research Institute (DDMRI) at the Nelson R Mandela School of Medicine of the University of KwaZulu-Natal in Durban, South Africa (Van Loggerenberg et al., 2008; Yirga et al., 2020b). Between August 2004 and May 2005, CAPRISA introduced a cohort study registering high-risk HIV-negative women to a follow-up study with an intense ongoing examination. Women infected with HIV were recruited into the CAPRISA 002 Acute Infection (AI) study and then followed up carefully to study disease progression and CD4/viral load evolution (Garrett et al., 2018; Mlisana et al., 2014; Moosa et al., 2018; Van Loggerenberg et al., 2008; Yirga et al., 2020a,b).

Once HIV-infected women were enrolled in CAPRISA's AI Phase II study, their CD4 count and viral load will be measured and assessed regularly. When their CD4 count ≤ 350 cells/mm^3 for more than two consecutive visits between six months or if they are with AIDS-defining illness (WHO clinical stage 3-5), they would be referred to a public government clinic for ARV treatment. However, according to the South African National Department of Health, these patients would only start HAART once their CD4 count is $\leq$ 200 cells/mm^3, until 2015. With effect from the 1st of January 2015, according to the National Department of Health, the criteria to start HIV patients on early initiation of ART is CD4 count ≤ 500 cells/mm^3 (Yirga et al., 2020b). HIV-infected women in Phase II-IV were followed up until they are started HAART. After that, they would be transitioned to Phase V and followed up for a minimum of five years, or eligible participants would be offered to join immediately into Phase V (Karim et al., 2017). After the five years of follow-up have been accomplished, participants would be offered an optional annual follow-up for up to fifteen extra years to patients who recurred in Phase V (Karim et al., 2017). Fig. 6.1 illustrates the screening and enrolment process of the study data set. One can find further detail on the study population's design, development, and procedures here (Garrett et al., 2018; Mlisana et al., 2014; Moosa et al., 2018; Van Loggerenberg et al., 2008; Karim et al., 2017).

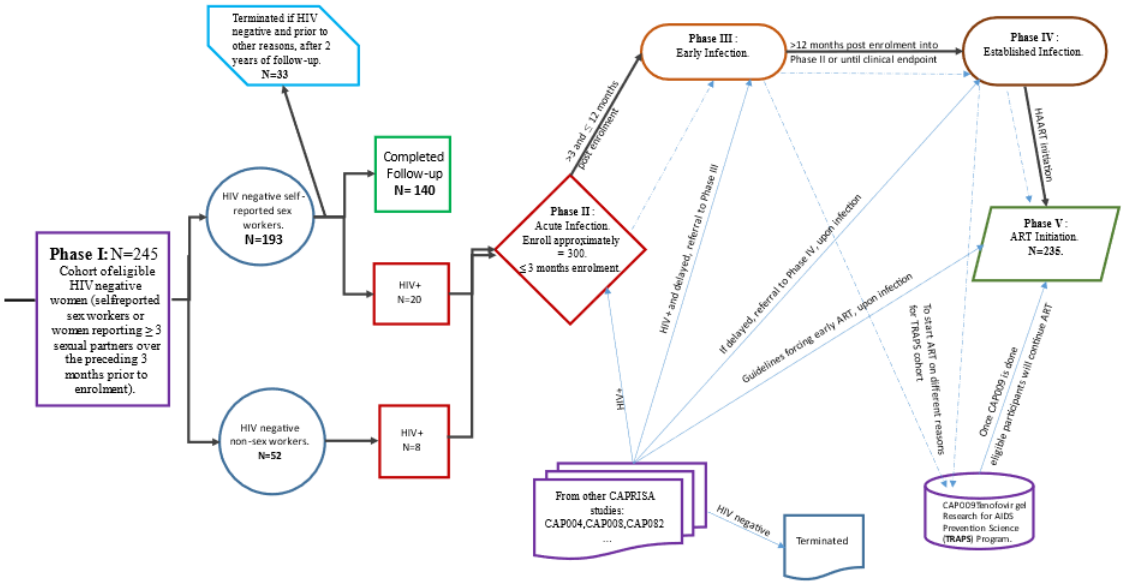


Figure 6.1: Diagrammatic overview of the CAPRISA 002 AI cohort study design

Geraci (2019) illustrated the full range of AQMM that is described above. The purpose of this analysis is to model the CD4 count of patients from KwaZulu-Natal, South Africa, as part of a comprehensive study of HIV/AIDS. The results of this study illustrate longitudinal CD4 counts among HIV-infected patients enrolled in the CAPRISA 002 AI study by employing an AQMM. The median age of our sample of 235 women was 25 years. Our sample consisted of 7019 measurements on 235 women from 18 to 59 years of age. There were multiple visits for all participants, ranging from 2 to 61, with a median of 29.

Tables 6.1 and 6.2 show the descriptive measures of the study variables. descriptive measures for the variables studied. Low (upper) quantiles are those where at least 25% (75%) of the observations are at or below it, or 75% (25%) are at or above it (Koenker, 2005a). In Table 6.1, it is shown that the median BMI for the participants was 26.84 (range 17.89 - 54.89). The median square root CD4 count and baseline viral load were 22.98 $cells/mm^3$ and 26600 copies, respectively. Of a total of 235 women, 105 (44.7%) lived around Vulindlela (rural area), and 130 (55.3%) lived around eThekwini (Durban, urban area) in KwaZulu-Natal, South Africa (see Table 2). The majority of the women, 182 (77.4%), were in a stable partnership, 224 (95.3%) completed secondary school (Table 6.2), and most of them (78.8%) were self-reported sex workers (Van Loggerenberg et al., 2008; Mlisana et al., 2014; Yirga et al., 2020b). Additional details are available here (Van Loggerenberg et al., 2008; Mlisana et al., 2014; Moosa et al., 2018; Garrett et al., 2018) concerning the CAPRISA 002 AI study. We analyze this data set intending to explain the different conditional distribution

of the CD4 count by considering two covariates enter as nonparametric additive effects: time and baseline BMI; as well as discrete, continuous, and categorical covariates enter in the model as parametric effects (see Tables 6.1 and 6.2). Fig. 6.2 shows observed square root transformed CD4 counts by treatment time and baseline BMI, respectively, for a total of 7019 observations. The nonlinear patterns, which connect the sample quantiles, are estimated conditionally on time and baseline BMI for six quantile levels. The curves (nonlinear patterns) suggest the requirement of some degree of smoothing (Fig. 6.2).

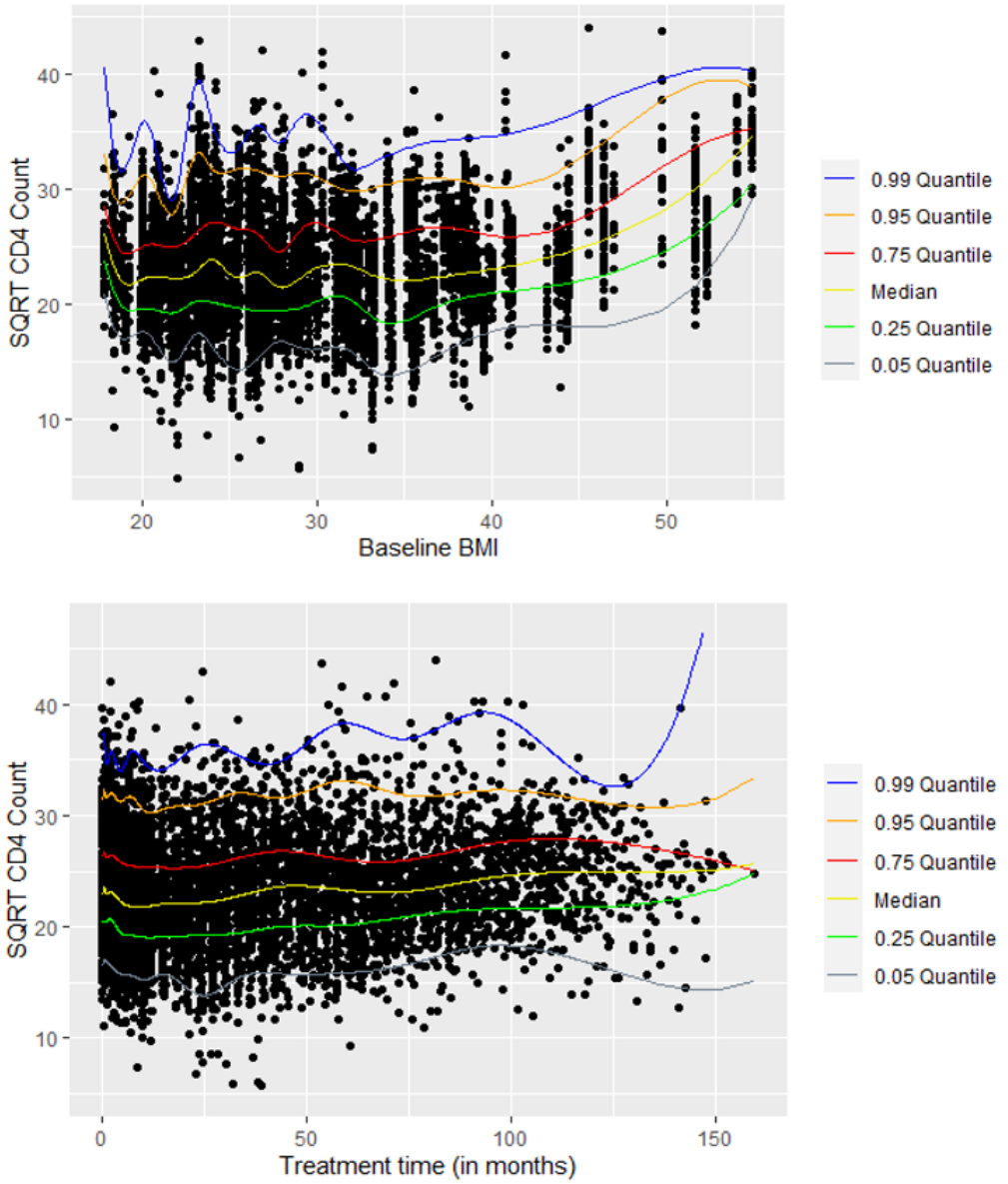


Figure 6.2: Observed CD4 counts (square root transformed) by time and baseline BMI across quantile levels

Following the AQMM of Geraci (2019), we used a transformed continuous form of the outcome (i.e., square root CD4 count) for fitting purposes. Thus, the proposed τ^{th} AQMM form of our study, using expression (6.7), can be specified as

$$\begin{aligned} Q^{*}_{y_{ij}|\boldsymbol{u}_i,\boldsymbol{x}_i,\boldsymbol{z}_i}(\tau) = \beta_{\tau,0} + \sum_{h=1}^{H_1} \mathcal{V}_{\tau,1}\, B_h^{(1)}(time_i) + \sum_{h=1}^{H_2} \mathcal{V}_{\tau,2}\, B_h^{(2)}(BMI_i) + \\ \beta_{\tau,1}\, ART_i + \beta_{\tau,2}\, VL_i + \beta_{\tau,3}\, residence_i + \beta_{\tau,4}\, education_i \\ + \beta_{\tau,5}\, partner_i + \beta_{\tau,6}\, age_i + u_{\tau,0} + u_{\tau,1}(time_i), \end{aligned} \tag{6.10}$$

where y_{ij} is the square root transformed form of the outcome ($\sqrt{\text{CD4 count}}$) at the j^{th} time point for the i^{th} subject, time is the time variable measured in months from the start of the study, BMI indicates the patient's baseline BMI, ART is the dichotomous HAART initiation (0 = pre-ART, 1 = post-ART), VL is patient's baseline viral load, the residence is patient's place of residence, education is the educational level of participants, partner indicates the number of sexual partners of the participant, age is participant age at enrolment, $u_{\tau,0}$ indicates the random intercept, and $u_{\tau,1}$ indicates the random slope. The symbol τ specifies the quantile of interest; we made the estimation at $\tau = 0.05, 0.25, 0.5, 0.75, 0.85, 0.95$, and 0.99 to get the complete picture of the effects.

Table 6.1: Descriptive statistics for non-categorical variables

	Descriptive measures						
Variables	**Mean**	**Median**	**Minimum**	**Maximum**	$Q_{0.25}$	$Q_{0.75}$	IQR
$\sqrt{CD4count}$ (cells/μ L)	23.26	22.98	5	44	20	26.19	6.19
Baseline VL (cells/mL)	130730.33	26600	1 (undetected)	5510000	5080	113000	107920
Age (Years)	27.15	25	18	59	22	30	8
Body Mass Index	28.98	26.84	17.89	54.89	23.33	32.96	9.63

Geraci (2019) employed the AQMM in the R package *lqmm* as an ad-on to fit additive quantile mixed models, and it is available from the author's GitHub platform (https://github.com/marco-geraci/aqmm). As the same as the smooth terms' specification in the R package *mgcv* (Wood, 2017), one can enter smooth terms continuous covariates within the *s* (smooth) function to control the model smoothness using splines when fitting AQMM (Geraci, 2019). Furthermore, the shrinkage smoothers obtained using the bs option inside the *s* command in the R package *mgcv* are constructed so that smooth terms can be penalized away altogether, not contribute to the model (Wood, 2017; Zuur et al., 2009). Thin plate smoother provides statistical and computational efficiency, stable optimal approximations (especially

Table 6.2: Baseline descriptive statistics for categorical variables

Variable	Total	Variable	Total
Number of women	235	**Number of sexual partners**	
Place of residence		No partner (*reference*)	43 (18.3%)
Rural (*reference*)	105 (44.7%)	Stable partner	182 (77.4%)
Urban	130 (55.3%)	Many partners	10 (4.3%)
Educational level			
Primary schools (*reference*)	11 (4.7%)		
Secondary schools	224 (95.3%)		

for large data sets), and can be constructed for smooths of more than one covariate at a time (Wood, 2003; Geraci, 2019). Thus, it was used as a shrinkage spline to fit the proposed model (6.10). The remaining parametric terms in the *aqmm* function (Geraci, 2019) are specified as the same as in other R linear mixed model fitting functions such as *lqmm ()* and *lme4 ()*. The output is separated into two parts: Parametric part that includes estimated fixed effects with their standard errors (SE), in parentheses, and significant mixed effect representation of smoothing splines (see Table 6.3). Since the smooth coefficients are mostly uninterpretable, we focus on their variances to evaluate the spline coefficients' penalty at various quantiles (see Table 6.4 and Table 7.8 in Appendix B). However, their estimated smoothed effects are depicted in Fig. 6.2. Table 6.4 also presents the estimated variance of the random effects from the fitted model (6.10).

According to Table 6.3, the age effect is positive and significant at the bottom, median, and at $\tau = 0.75$ quantile levels (see also Supplementary material 1). On the other hand, the effect of education on square root of CD4 count does not seems to be significant across all quantiles after the patient had been initiated on HAART. The square root of CD4 count across all quantiles is affected by post-HAART initiation as expected. A significant positive effect of HAART initiation on CD4 cell counts is observed at the median quantile and becomes roughly constant at higher quantiles (see Table 6.3 and Table 7.8 in Appendix B). In addition, patients with stable sexual partners showed significant improvements in their CD4 cell count across all quantiles. The CD4 cell count is significantly lowered in patients who have many sexual partners, especially at the bottom ($\tau = 0.05$) and at the top ($\tau = 0.95, 0.99$) quantiles (Table 6.3).

Furthermore, we found a clear indication, at the bottom ($\tau = 0.05$) and more extreme

Table 6.3: Parameter estimates followed by results of the smoothing terms from the AQMM for the CAPRISA 002 AI study data across different quantiles

Fixed effects	$\hat{Q}_{0.05}$ **(SE)**	$\hat{Q}_{0.25}$ **(SE)**	$\hat{Q}_{0.5}$ **(SE)**	$\hat{Q}_{0.75}$ **(SE)**	$\hat{Q}_{0.95}$ **(SE)**
Intercept	16.004 (0.6634) ***	19.647 (0.4749)***	21.204 (0.5340) ***	24.167 (1.0536)***	29.379 (0.6324) ***
Age	0.0398 (0.0156) **	0.0209 (0.0114) .	0.0418 (0.0052) ***	0.0331 (0.0078)***	0.0203 (0.0178)
Secondary school	-0.4491 (0.5731)	-0.4734 (0.4101)	-0.0165 (0.6619)	0.0385 (1.0677)	0.8323 (0.5574)
Post HAART	0.7430 (0.0879) ***	1.5296 (0.0598)***	1.5968 (0.0402) ***	1.5292 (0.0546)***	1.7007 (0.1322) ***
Baseline VL	-3.83e-06 (8.42e-07) ***	-2.09e-06 (2.69e-07)***	-1.79e-06 (2.41e-07) ***	-1.57e-06 (1.60e-07)***	-1.70e-06 (2.21e-07) ***
Urban	-0.50002 (0.1668) **	0.2499 (0.0545)***	0.0998 (0.0334) **	0.1275 (0.1436)	-0.8846 (0.2216) ***
Stable partner	0.6135 (0.1655) ***	0.3046 (0.1549) .	0.5424 (0.1140) ***	0.4907 (0.1594)**	0.6339 (0.2960) *
Many partners	-2.2771 (0.2707) ***	-0.7858 (0.2589)**	-0.8432 (0.1091) ***	-1.1719 (0.2569)***	-3.6497 (0.4451) ***
Results of the smooth terms					
s (Time)	-2.5075 (0.5426) ***	-2.3766 (0.5549)***	-2.1985 (0.4735) ***	-2.2829 (0.4999)***	-2.3324 (0.4373) ***
s (Baseline BMI)	5.4382 (1.0786) ***	5.6868 (1.1094)***	5.5767 (1.3014) ***	5.7904 (1.2077)***	5.2604 (1.0753) ***

- Significance codes: 0 '***', 0.001 '**', 0.01 '*', 0.05 '.', 0.1 ' ', 1.
- The reference categories are given in Table 6.2.

quantiles ($\tau = 0.85, 0.95, 0.99$), that there is a significant negative effect of patients who were residing around the urban area, on their CD4 cell count (see Table 6.3 and Table 7.8 in Appendix B). Table 6.3 also shows that the negative effect of baseline viral load on the CD4 cell count is higher at the lower quantiles than at the median and higher quantiles (see, also, Table 7.8 in Appendix B). In addition, R package *aqmm ()* sample outputs using CAPRISA 002 AI study data at $\tau = 0.25, 0.75, 0.85$, and 0.99 can be found in Table 7.8 in Appendix B.

Table 6.4: Estimated variance of the random effects and smooth terms from the AQMM for the CAPRISA 002 AI study data

	$\hat{Q}_{0.05}$	$\hat{Q}_{0.25}$	$\hat{Q}_{0.5}$	$\hat{Q}_{0.75}$	$\hat{Q}_{0.85}$	$\hat{Q}_{0.95}$	$\hat{Q}_{0.95}$
Variance of the random effects							
$\hat{\sigma}_0$ (Intercept)	2.748e-02	8.687e-01	3.543e-02	2.453e-01	3.454e-01	4.675e-02	3.326e-03
$\hat{\sigma}_0$ (Time)	8.104e-18	1.929e-16	3.328e-17	5.451e-17	7.671e-17	1.044e-17	2.963e-18
Variance of the smooth terms							
$\hat{\phi}_{Time}$	8.796	28.94	36.74	30.28	21.92	10.13	2.669
$\hat{\phi}_{BaselineBMI}$	1876.501	6463.83	7823.81	6290.32	4979.39	2183.69	576.902

The variance of the first smooth term ($\hat{\phi}_{Time}$) indicates a stronger penalty on the spline coefficients at $\tau = 0.25, 0.5, 0.75, 0.85$ quantiles than at the bottom and at the top quantiles (Table 6.4). Similarly, the variance of the second smoother ($\hat{\phi}_{BaselineBMI}$) shows a strong penalty on the spline coefficients at $\tau = 0.25, 0.5, 0.75, 0.85$ quantiles than at the bottom and at more extreme quantiles. Table 4 shows that the random effects' variances have roughly constant variability of subject linear trends across the fitted quantiles (see, also, Table 7.8 in Appendix B).

Based on the seven fitted quantile levels ($\tau = 0.05, 0.25, 0.5, 0.75, 0.85, 0.95, 0.99$), Fig. 6.3 depicts the two estimated smoothed covariate effects on patients' CD4 counts. Patients enrolled in the CAPRISA 002 AI study exhibit nonlinear time effects on CD4 counts that are prominent at all quantile levels. As the quantile increases, its effect becomes stronger. However, it is after several treatment visits that such progress towards higher CD4 counts occurs. Consequently, the progression is slow until about 50 months, then it increases steadily thereafter across all quantile levels (Fig. 6.3).

Furthermore, overall fit quantile levels, the significant smoothed baseline BMI effect on patients' CD4 counts is roughly constant for patients with a baseline BMI of about 40 but gradually improves from there. Because of this, patients with low BMI need to be monitored carefully before and after HAART initiation. Despite this, physicians should not ignore patients with high BMI. According to our studies and other findings, a plausible explanation may be that BMI may affect drug metabolism and, thus, the progress of HAART and its immunological responses (Palermo et al., 2011; Li et al., 2019; Yirga et al., 2020a,b). Moreover, higher levels of BMI have a more significant effect than lower levels (Fig. 6.3).

Some of the codes that were used for this section can be found here (Code 7.5 in the Appendix A).

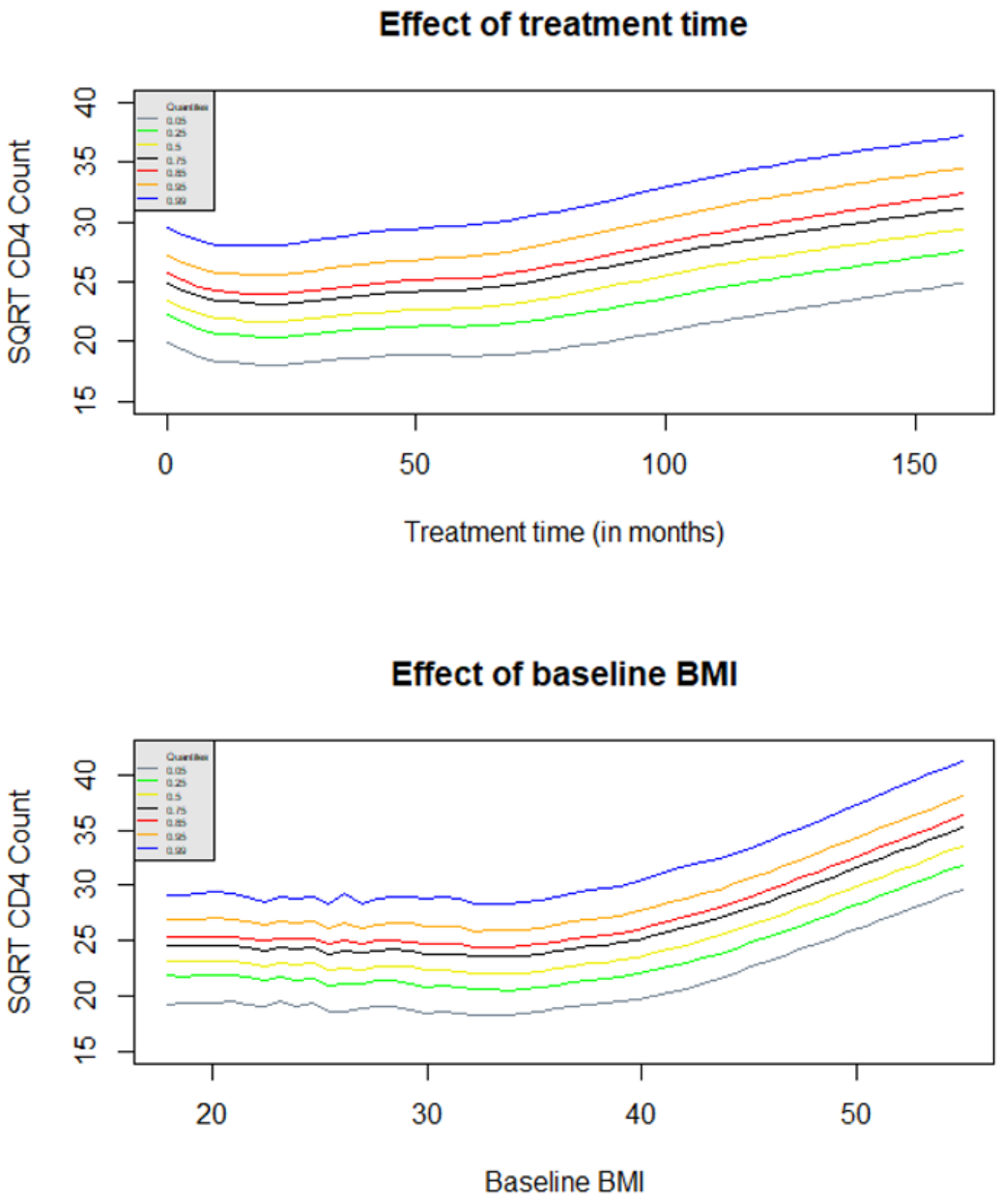


Figure 6.3: Predicted smoothed covariate effects on the square root CD4 count of HIV-infected patients recurred in the CAPRISA 002 AI study at various quantiles using AQMM

6.6 Summary

In this chapter, we discussed the additive quantile model and also considered additive quantile mixed models of Geraci (2019) to capture the parametric and nonparametric covariate effects on the longitudinal CD4 count of HIV-infected patients across various quantile levels. It turns out that this recently developed model can be used to obtain robust results, not only at the central location of the longitudinal outcome that may not be the best location to characterize the data but also at different locations of the conditional distribution that communicates an inclusive and more complete picture of the parametric as well as the nonparametric covariate effects.

A series of AQMM at $\tau = 0.05, 0.25, 0.5, 0.75, 0.85, 0.95$, and 0.99 were performed, and the results were discussed. According to the results, patients' CD4 count is markedly increased after HAART initiation, and their baseline viral load shows a negative effect on the progression of their CD4 count over time, as we would have expected. All fitted quantiles of the response variable were affected by a significant nonlinear relationship between time and baseline BMI. Study results suggest that, across all fitted quantile levels, the patient's education level does not significantly influence the progression of CD4 counts over time. All but the most extreme quantiles of HIV-positive patients showed a significant difference in the CD4 count regardless of their age. In addition, CD4 cell recovery was found to be significant across all quantiles among patients with a stable sexual partner. Contrary to this, HIV-infected patients with many sexual partners during the treatment period showed a negative effect on CD4 cell count across all fitted quantile levels.

As we expected, the patient's CD4 count significantly increased after HAART was initiated, and their baseline viral load also showed a significant negative effect on the patient's CD4 count over time. Baseline BMI and time were also significant nonlinear effects in our analysis. Further, patients with higher BMIs at baseline have improved CD4 cell count over time after treatment. Despite this, higher BMI patients should not be ignored clinically. This study instead suggests that BMI can influence drug metabolism and, consequently, the immunological responses to HAART. According to the nonlinear time effect, patients' CD4 counts are not increasing rapidly over time. The growth starts after multiple treatment visits. Hence, the study suggests that HIV patients who are not clinically and immunologically stable on HAART could experience increased risks if exposed to COVID-19, especially if they are not on HAART immediately after HIV exposure.

One can estimate the covariate effects over the grid $\tau \in (0, 1)$ as per the analysis

aspects. An investigator, however, should be cautious when using AQMM since the method needs some adjustment to control the estimation algorithm and demands more computing time to estimate the random effects (Geraci, 2019). For instance, for this study, it took 2 – 3 hours to fit the proposed model (6.10) at a single τ as like Geraci (2019). To overcome this computational burden, Geraci (2019) suggested the necessity of further improvement to the AQMM. As the studied data is an on-going study, there is a plan to extend AQMM application to gene expression studies (machine learning framework) in future work since it produces satisfactory results.

Chapter 7

Discussion and Conclusion

Parametric and nonparametric driven models for longitudinal forms of data analysis more closely resemble mixed-effects models, and their extensions to the exponential family of distribution, quantile regression-based, and additive-based models have been the focus of this dissertation. These methodologies, with detailed discussion, have been employed on the subset of the Centre for the AIDS Programme of Research in South Africa's (CAPRISA) data set. For a detailed discussion of types of methodologies that are employed in this thesis, see Diggle et al. (2002), Pinheiro & Bates (2006), Fitzmaurice et al. (2012), Der & Everitt (2012), Gbur et al. (2012), Zuur et al. (2009), Molenberghs & Verbeke (2006), Liu (2015), Demidenko (2013), Hilbe (2011, 2014), Koenker (2005a), Yu & Moyeed (2001), Yu & Zhang (2005), Machado & Silva (2005), Geraci & Bottai (2014), Galarza et al. (2015), Xiang (2001), Hastie (2017), Wood (2017), Stroup (2012), Fenske et al. (2011, 2013), Geraci (2019), and the references therein.

This thesis began with reviewing and applying a mixed-effects model to visualize and understand the longitudinal data set we used for analysis. In mixed-effects models, we can account for random variation if one group has more variability than another or if we want to analyze correlations over time. Unlike linear models, mixed-effects models can be used to deal with multi-level, clusters, dependencies in the data, and missing values. It also works well with unbalanced designs. Thus, mixed-effects models are suitable for both complete and incomplete data. Therefore, all of these make the mixed-effects model a powerful technique to study longitudinal data. In Chapter 2, we have also discussed some of the terminologies in mixed-effects models, such as the difference between fixed and random effects, different types of random-effects models, and the ML and REML estimation techniques.

The appropriate selection of random effects using the REML estimation technique resulted in the random intercept and slopes model, which incorporated the intercept, time, and square root of time, as the best model. A comparison of covariance structures was made to have a valid inference about the mean structure. The results of Chapter 2 suggested that the UN covariance structure was the best structure for the fitted model. The result also confirmed that patients' average CD4 count significantly increased after the patient had been initiated on ART. It was found that patients with higher BMI at the baseline showed a considerably increasing number of CD4 counts after being initiated on HAART. Moreover, the result confirmed that patients with higher viral load before the patient being initiated on HAART experience a significant adverse effect on the prognosis of patients CD4 count.

Influence and model diagnostics were also conducted in Chapter 2. Since our data set is identified as unequally measured longitudinal data, a comparison of the three commonly used spatial covariance structures: spatial exponential structure (SP(EXP)), spatial spherical structure (SP(SPH)), and spatial Gaussian structure (SP(GAU)), were conducted to measure the actual distance or variation among observations as well as to account for spatial variability (heterogeneity). It was found that the SP(EXP) is the best spatial covariance model. Note that the "spatial" indicated just the name of the correlation structure that uses spatial in it.

Following the mixed-effects model that was conducted concentrating on the transformed continuous normalized response variable (square root of CD4 count) was extended to exponential families of distribution in Chapter 3 to analyze the non-normal, over-dispersed, and non-transformed longitudinal count data. The generalized linear mixed-effects model (GLMM) allows for both normal and non-normally distributed response variables; that is where it gets the "generalized" term form. It also enables predictor variables to be either fixed and/or random (subject-specific) effects, which is the "mixed" part of the model. GLMM incorporates random effects to model the correlation between observations. GLMMs are also one of the most valuable methods when the main scientific objective is to make inferences about subject-specific effects. For these reasons, GLMMs cover a wide variety of models, from simple linear regression to complex multi-level models for non-normal and normal longitudinal forms of data.

The properties of mixed-effects models that include random effects and generalized linear models that handle non-normal data by letting the errors take on exponential family of distribution can be combined to model such instances in a GLMM framework effectively. Inference on the GLMM uses the same basic ideas as the

conventional mixed-effects models and the generalized linear models. However, in GLMMs, the mean of the response and the predictors are modeled through the *link* functions.

For count data, over-dispersion is potentially one of the leading modeling topics in applied statistics. If it is not taken into consideration, it may lead to an inadequate result. Thus, a reasonable GLMM approach that manages over-dispersed longitudinal count data is used in this chapter. Since a Poisson process is mainly used as an initial point for modeling the stochastic differences of count data with the canonical link being the log, the Poisson mixed-effects model was first employed in Chapter 3. However, due to its lack of realistic properties, such as the restriction that the mean and variance are equal, the Poisson mixed-effects model is replaced by the negative binomial mixed-effects model. Thus, the later model showed appropriate properties and out-performed the PMM model to manage the over-dispersion of our longitudinal count data.

The parameter estimates based on the NBMM are not exceptionally different from those based on PMM. However, as mentioned above, the PMM approach leads to inadequate results when over-dispersion is present. Some of the available methods to estimate the parameters in GLMMs were discussed. Our preference goes to the Laplace approximation due to the fewer limitations than the Adaptive quadrature and its accuracy, fast and plausibility to use the likelihood as well as the information criteria.

Little's missing completely at random (MCAR) test was used to check whether the missing values in our data set are MCAR or not. It was found that the missing data in the study variables of interest were not MCAR. Therefore, multiple imputation techniques were used to handle missing values in the data set to validate parameter estimates from the complete data set using NBMM. However, due to the relatively low amount of missing data in the analysis variables, we did not find major differences. In both cases, analysis with missing values, and multiple imputation analysis, covariates that were found to be significantly affecting CD4 count of the patient were similar, and their respective parameter estimates are more close to each other. Therefore, missing data analysis was not the scope of the study in any of the other chapters in the thesis. It was found that the effect of treatment time, baseline BMI, post-ART initiation, baseline viral load, and the number of sexual partners on the patient's CD4 count, as highly significant factors in Chapter 3.

Further, in Chapter 4, a quantile mixed-effects approach was proposed to detect the

heterogeneous effects of covariates at the conditional distribution of the response. Quantile regression offers an invaluable to discern effects that would be missed by other conventional regression models, which analyze the sole conditional mean. Estimated effects of mixed-effects models and GLMMs are formulated on the response variable through mean-based regression. However, this centrality-based inferential system cannot represent the entire distribution of the outcome and, in some cases, may not be the best location to characterize the longitudinal data. While regression for medians can be seen as more robust than regressions to model the mean value, quantile regression, a generalization of median regression, enables more fully to explore the data by modeling the conditional quantile at low or high quantiles such as the 5^{th} or 95^{th} percentiles of the response distribution. For these reasons, quantile regression emerged as an effective analytic technique in numerous study areas of science.

Despite the fact quantile regression was primarily established in a univariate setting, the considerable amount of longitudinal data recently dictates its extensions towards a mixed-effects modeling system (Liu & Bottai, 2009; Geraci & Bottai, 2007; Galarza et al., 2017). Quantile mixed-effects model has become practical for longitudinal data analysis due to the recent computational advances and ready availability of efficient linear programming algorithms. Thus, its application has received increasing consideration in wide-ranging areas of study (Lachos et al., 2015; Fu & Wang, 2012; Geraci & Bottai, 2014; Galarza et al., 2015). Thus, the QR-LMM (Galarza et al., 2017) for our longitudinal data is applied in this chapter. The QR-LMM concept is similar to that of the conventional quantile regression for independent data. However, there are differences in the estimation due to the existence of random effects in the QR-LMM. As in the linear model, the estimation of the regression parameter β_τ can be processed using the ML estimator problem by assuming an ALD unit error model. At the same time, the random effects in the QR-LMM need to be predicted. Thus, the QR-LMM estimator combines the ML estimator of β_τ and the random effect predictor.

QR-LMM is a likelihood-based function that adopts an ALD for the error term, as mentioned above, in which multiple random effects can be incorporated into the model to account for the dependence among the longitudinal data. This method uses the SAEM algorithm for determining exact ML estimates of the covariates effects and variance-covariance elements across a set of quantiles. We further applied this methodology to a subset of CAPRISA data to justify how the procedure developed can be used to obtain robust parameter estimates when the interest is to get the estimation at different locations of the conditional distribution, which then brings a

comprehensive and more complete picture of the effects. A series of QR-LMM over the grid $\tau = \{0.05, 0.25, 0.5, 0.75, 0.85, 0.95\}$ were estimated, and the results were discussed. It was found that treatment time (measured in a month), ART-initiation, baseline BMI, and baseline viral load, were significant factors of the patient's CD4 count across all quantile levels.

Since there is always a complex form of relationships between the outcome variable and the predictors, unknown covariates' functional form, and inflexibility in parametric models, the generalized additive mixed-effects approach is conducted in Chapter 5. The objective of the chapter is to let the data decide the relationship between variables. Additive models and, more generally, generalized additive models are a generalization of nonparametric regression models in which they deal with non-linearity in covariates that are not the main interest in a study and adjust for those effects appropriately while still retaining much of their interpretability. GAMs enable the mean of the outcome to rely on an additive predictor via a nonlinear link function. The GAMs consist of an additivity assumption that enables relatively many nonparametric relationships to be examined simultaneously and the distributional flexibility of GLMs.

GAMs differ from conventional linear regression methods by allowing the so-called *smooth terms* alongside parametric representations. The coefficients for the individual *basis functions* (knots) contained in a GAM smooth term are estimated in such a way that the resulting curve has a controlled degree of wiggliness determined by the smoothing parameter. The *estimated degree of freedom* (edf) is used to test the significance of the smoother term in GAMs; as a result, analysis results of GAMs always come with these edf values. However, features of the smoothing function in GAM are often examined by graphical visual inspection.

The generalized additive mixed model, a mixed-effects version of GAM, is another powerful method that deals with non-normally distributed outcome, non-linear trajectories in the longitudinal data using nonparametric regression, and accounts for within-individual correlation (hierarchical structure of the data) by incorporating random effects (Lin & Zhang, 1999). GAMMs extend the GLMMs by allowing continuous predictors to have a smooth functional impact on the mean response (Lin & Zhang, 1999). GAMM can be fitted using the *gamm* or *gamm4* function from the *mgcv* package in *R*; the function *gam.check ()* can be used to produce residual plots (model-checking).

In Chapter 5, we employed an additive negative binomial mixed-effects model, an

example of a GAMM that accommodate over-dispersion of the data, as an extension to Section 3.6 and 3.7 to analyze the multiple repeated measures of patients' CD4 cell count, as a function of age, baseline BMI, and treatment time non-parametrically, and baseline viral load, HAART initiation, level of education, place of residence, and the number of sexual partners parametrically. The results of the analysis gave us more insights to look into the functional relationship between the response variable and the covariates. The study confirmed that the linear effect of HAART initiation and baseline viral load have a significant positive and substantial adverse effect, respectively, on the progression of patients' CD4 cell count over time. Furthermore, the analysis revealed that the relationship between patients' CD4 count and each of the nonparametric terms (age, baseline BMI, and time) could be better explained by a nonlinear relationship.

In Chapter 6, we examined an extended form of additive mixed-effects model to quantile-based regression. A comprehensive analysis of a variety of different covariate (parametric and nonparametric) effects was observed, not only at the mean level of the longitudinal outcome, which is not necessarily the best place to characterize the data but also across different locations of the conditional distribution. Additive quantile mixed model (Geraci, 2019), is a recently developed model that gained popularity as a general method for analysis of longitudinal data.

As previously discussed, quantiles, especially the median, are important to understand and play a fundamental role in statistics. By definition, mean-based analysis average out stronger and weaker effects. The averaging may even cancel out symmetric effects of some magnitudes but opposite signs on the tail of the distribution (Geraci, 2019). Quantile-based regression emerged as an effective analytic technique in numerous study areas of science due to its flexibility to make inferences focused on features of the response beyond its central tendency. It is especially appropriate for the parameters' heterogeneous effect as it yields inferences that can be legitimate irrespective of the true underlying distribution. Quantile regression techniques look further into the data, get more information, and become an essential statistical tool for addressing numerous research questions.

As part of this thesis, we have reviewed and applied few aspects of quantile-based models. More effective work can also be done based on quantile regression and its extension by other researchers. Apart from its wide-ranging scientific areas of applications, Koenker (2005b) quantile regression has also been extended to various statistical techniques: additive quantile regression, and additive quantile mixed model, to say the least. Additive quantile regression combines quantile regression with an

additive predictor that consists of smooth non-linear effects of continuous covariates thereby enables a variety of covariate effects to be flexibly modeled (Fenske et al., 2011). Furthermore, like the conventional additive model, the additive quantile regression model does not require a predetermined functional fit but instead determines the best fit from the data.

As an extension to AMMs, the additive quantile mixed model consider the effect of linear and non-linear, as well as random effects across various quantiles of the conditional distribution; thus, it has an advantage over the AMM in non-linear and heteroscedastic cases (Geraci, 2019). The GAMLSS approach of Rigby & Stasinopoulos (2005) that uses parametric methods based on flexible distributions can be considered as an equivalent alternative to fit the linear and non-linear trajectories of the model at different quantiles. But, they do not provide easy coefficient interpretation of the *quantile treatment effect* of the covariates (Fenske et al., 2013; Geraci, 2019). Geraci (2019)'s AQMM approach that aims at the conditional quantiles of the dependent variable without assuming any distribution for the error term, estimate the level of smoothing of the nonparametric terms automatically from the data, and provides convenient regression coefficient interpretation as like the conventional QR model, also has an advantage compared to GAMLSS and that of Fenske et al. (2013)'s additive fixed effects quantile regression model for longitudinal data. In addition, AQMM has unique features compared to other alternative or additive-based approaches; it provides the mixed-effects representation of smoothing splines that leads to automatic smoothing selection and able to model the variance-covariance matrix of the random effects (Geraci, 2019).

The measure of the accuracy of an estimate such as standard error and confidence interval in AQMM is facilitated by the bootstrap method, which is a general resampling procedure, adopted in Kleiner et al. (2014) as *bag of little bootstraps* (BLB) approach (Geraci, 2019). The BLB approach includes features of both the bootstrap and subsampling to obtain a robust, computationally efficient means of assessing the quality of estimators (Kleiner et al., 2014). Confidence intervals resulted from the bootstrap procedures implemented in quantile regression models have shown asymptotically valid coverage probabilities (Hahn, 1995). Geraci & Bottai (2014) also worked on bootstrapping confidence intervals in their linear quantile mixed models, and they showed that the results have good coverage probabilities (Geraci, 2019). Efron (1992) introduced the bootstrap technique as a computer-based method for estimating the distributions of statistics using the observations of the sample. Bootstrapping does not require any distributional assumptions, provides more accurate inference even when the sample size is small, and can be applied to statistical meth-

ods with sampling distributions that are difficult to drive asymptotically. The basic idea of the bootstrapping method is that it selects an observation or random sample from the original data, with replacement, to obtain an ideal estimate of the sampling distribution of interest (e.g., variance, confidence intervals, prediction error) (Efron, 1992; Efron & Tibshirani, 1994).

The results from a series of AQMM at $\tau = 0.05, 0.25, 0.5, 0.75, 0.85, 0.95,$ and 0.99 were reported. Interestingly, the study revealed that except at the more extreme quantile levels, patient's age was found to have a significant linear effect on the progression of their CD4 count across the fitted quantiles. It was also observed that significant CD4 cell recovery in response to patients with a stable sexual partner across all fitted quantiles. In contrast, HIV-infected patients with many sexual partners during the treatment period showed a significant adverse effect on CD4 cells recovery across all fitted quantiles. Furthermore, as like Chapter 5 study result, the analysis of AQMM also confirmed that time and baseline BMI were found to have a significant nonlinear effect on the patients' CD4 count across all fitted quantile levels. In line with all the previous chapters, the result of the AQMM showed, across all fitted quantiles, the progression of patients' CD4 count is significantly increased after HAART initiation, and patients' baseline viral load showed a significant adverse effect on the recovery of their CD4 count over time.

In conclusion, in this dissertation, we reviewed and applied different parsimonious longitudinal data-based modeling approaches. That includes mixed-effects model, generalized linear mixed-effects models such as Poisson-based and negative binomial-based models, quantile mixed-effects model, generalized additive mixed models such as additive negative binomial mixed-effects model, and additive quantile mixed-effects model to analyze the number of CD4 cells measured repeatedly in HIV-infected patients enrolled in the CAPRISA study. In addition, several important data features such as within-subject correlation, heterogeneity between subjects, variation at lower and higher levels of the design structure (fixed and random effects), missing value analysis, identification of influential observations, models diagnosis, covariance and spatial auto-correlation structures, over-dispersion, average effect across the whole population and individual level, both non-Gaussian and transformed Gaussian longitudinal forms of data, continuous and count forms of outcomes, parametric and nonparametric covariate effects, and the association between outcome and covariates across various quantile levels were also considered.

Although this dissertation research is motivated by the CAPRISA study data set, the novel methodologies employed in this thesis have broader applications and flexibil-

ity for investigators to explore further with their longitudinal forms of data. With the growing recognition of the quantile-based regression model, we may also extend our quantile-based model applications to other vital developments. The methodology has been developed to various statistical methods such as quantile-based survival models for longitudinal data, machine learning frameworks, and other critical areas.

Survival data analysis is part of statistical methods for data analysis for which the outcome variable of interest is time until an event occurs (Kleinbaum et al., 2012). When the event is more than one, the study is called competing risk analysis. The idea of competing risks is that everyone in the study or real life is subject to several hazards that cause an event or experience more than one type of a particular event (competing events). What is typical about survival analysis are two things: usually, survival time is not normally distributed, it is very often right-skewed, and survival times are incomplete either due to censoring or truncation. Because of these two reasons, survival data cannot be analyzed by standard statistical procedures such as linear regression. Compared to the famous Cox models, quantile-based survival analysis relaxes the proportional hazard assumptions, links the entire distribution of an outcome to the covariates of interest, and provides considerably more flexibility to explore the heterogeneous effects of covariates for a non-homogeneous population (Koenker & Geling, 2001; Hong et al., 2019).

Most machine learning methods such as Random Forest, Neural Network, Kernel functions, Support Vector Machines, and Gradient Boosting provide mean-based prediction intervals and perform better when the outcome of interest follows a standard (Gaussian or uniform) probability distribution. However, when the distribution of the target variable is heteroscedastic and when we are interested in the prediction intervals at various points, it is recommended to use quantile-based machine learning approaches. These shall be the subject of future works.

References

Abramowitz, M., & Stegun, I. A. (1948). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, vol. 55. US Government Printing Office.

Agresti, A. (2003). *Categorical data analysis*. John Wiley & Sons.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Allison, P. D. (2001). *Missing data*. Sage Publications.

amfAR (2015). The Foundation for AIDS Research. Statistics: Women and HIV/AIDS. Accessed: 2019-10-12.
URL https://bit.ly/3yKoLoN

Ayele, D. G., Zewotir, T. T., & Mwambi, H. G. (2014). Semiparametric models for malaria rapid diagnosis test result. *BMC Public Health*, *14*(1), 1–10.

Baayen, H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, *94*, 206–234.

Belsley, D. A., Kuh, E., & Welsch, R. E. (2005). *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons.

Berglund, P., & Heeringa, S. G. (2014). *Multiple imputation of missing data using SAS*. SAS Institute.

Berhane, K., & Tibshirani, R. J. (1998). Generalized additive models for longitudinal data. *Canadian Journal of Statistics*, *26*(4), 517–535.

Blankenberg, S., Salomaa, V., Makarova, N., Ojeda, F., Wild, P., Lackner, K. J., Jørgensen, T., Thorand, B., Peters, A., Nauck, M., et al. (2016). Troponin I and cardiovascular risk prediction in the general population: The BiomarCaRE consortium. *European Heart Journal*, *37*(30), 2428–2437.

Borgoni, R. (2011). A quantile regression approach to evaluate factors influencing residential indoor radon concentration. *Environmental Modeling & Assessment*, *16*(3), 239–250.

Boswell, M. (1970). *Chance mechanisms generating the negative binomial distribution*. Pennsylvania State, University Press.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345–370.

Breiman, L., & Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, *80*(391), 580–598.

Breslow, N. E. (1984). Extra-Poisson variation in log-linear models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *33*(1), 38–44.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*(421), 9–25.

Brown, H., & Prescott, R. (2014). *Applied mixed models in medicine*. John Wiley & Sons.

Buchinsky, M. (1998). Recent advances in quantile regression models: A practical guideline for empirical research. *Journal of Human Resources*, (pp. 88–126).

Bücker, M. J. D., & Hogan, J. W. (2011). Missing data in longitudinal studies. *Statistical Papers*, *52*(2), 501.

Cade, B. S., & Noon, B. R. (2003). A gentle introduction to quantile regression for Ecologists. *Frontiers in Ecology and the Environment*, *1*(8), 412–420.

Cameron, A. C., & Trivedi, P. K. (1986). Econometric models based on count data. Comparisons and applications of some estimators and tests. *Journal of Applied Econometrics*, *1*(1), 29–53.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications*. Cambridge University Press.

Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data*, vol. 53. Cambridge University Press.

Cameron, A. C., Trivedi, P. K., et al. (2009). *Microeconometrics using Stata*, vol. 5. Stata Press College Station, TX.

Casella, G., & Berger, R. L. (2002). Statistical Inference. *Duxbury. Pacific Grove, CA*.

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, *46*(3), 167–174.

Chambers, J. M. (1991). *Statistical models in S*. CRC Press, Inc.

Chaudhuri, P. (1991). Global nonparametric estimation of conditional quantile functions and their derivatives. *Journal of Multivariate Analysis*, *39*(2), 246–269.

Chen, C. (2000). Generalized additive mixed models. *Communications in Statistics-Theory and Methods*, *29*(5-6), 1257–1271.

Chen, C. (2005). Growth Charts of Body Mass Index (BMI) With Quantile Regression. *AMCS*, *5*, 114–20.

Christensen, R. (1991). *Linear models for multivariate, time series, and spatial data*. Springer Science & Business Media.

Christensen, R., Pearson, L. M., & Johnson, W. (1992). Case-deletion diagnostics for mixed models. *Technometrics*, *34*(1), 38–45.

Chunying, Z. (2011). A quantile regression analysis on the relations between foreign direct investment and technological innovation in China. In *2011 International Conference of Information Technology, Computer Engineering and Management Sciences*, vol. 4, (pp. 38–41). IEEE.

Cleveland, W. S., & Loader, C. (1996). Smoothing by local regression: Principles and methods. In *Statistical theory and computational aspects of smoothing*, (pp. 10–49). Springer.

Cohen, M. S., Shaw, G. M., McMichael, A. J., & Haynes, B. F. (2011). Acute HIV-1 Infection. *New England Journal of Medicine*, *364*(20), 1943–1954.

Cook, B. L., & Manning, W. G. (2009). Measuring racial/ethnic disparities across the distribution of health care expenditures. *Health Services Research*, *44*(5p1), 1603–1621.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, *19*(1), 15–18.

Cook, R. D. (1979). Influential observations in linear regression. *Journal of the American Statistical Association*, *74*(365), 169–174.

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman and Hall.

Craig, S. G., & Ng, P. T. (2001). Using quantile smoothing splines to identify employment subcenters in a multicentric urban area. *Journal of Urban Economics*, *49*(1), 100–120.

Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.

Davidian, M., & Giltinan, D. M. (2003). Nonlinear models for repeated measurement data: an overview and update. *Journal of Agricultural, Biological, and Environmental Statistics*, *8*(4), 387–419.

Davino, C., Furno, M., & Vistocco, D. (2013). *Quantile regression: Theory and applications*. John Wiley & Sons.

Dean, C., Lawless, J., & Willmot, G. (1989). A mixed Poisson–Inverse-Gaussian regression model. *Canadian Journal of Statistics*, *17*(2), 171–181.

Delyon, B., Lavielle, M., & Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics*, (pp. 94–128).

Demidenko, E. (2013). *Mixed models: Theory and applications with R*. John Wiley & Sons.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Der, G., & Everitt, B. S. (2012). *Applied medical statistics using SAS*. CRC Press.

Diggle, P., Diggle, P. J., Heagerty, P., Liang, K.-Y., Heagerty, P. J., Zeger, S., et al. (2002). *Analysis of longitudinal data*. Oxford University Press.

Dobson, A. J., & Barnett, A. (2008). *An introduction to generalized linear models*. CRC Press.

Dobson, A. J., & Barnett, A. G. (2018). *An introduction to generalized linear models*. CRC Press.

Duchateau, L., Janssen, P., & Rowlands, J. (1998). *Linear mixed models. An introduction with applications in veterinary research*. ILRI (aka ILCA and ILRAD).

Dufouil, C., Brayne, C., & Clayton, D. (2004). Analysis of longitudinal studies with death and drop-out: A case study. *Statistics in Medicine*, *23*(14), 2215–2226.

Durbán, M., Harezlak, J., Wand, M., & Carroll, R. (2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine*, *24*(8), 1153–1167.

Efron, B. (1992). Bootstrap methods: Another look at the jackknife. In *Breakthroughs in statistics*, (pp. 569–593). Springer.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC Press.

Ellerbe, C. N., Gebregziabher, M., Korte, J. E., Mauldin, J., & Hunt, K. J. (2013). Quantifying the impact of gestational diabetes mellitus, maternal weight and race on birthweight via quantile regression. *PLOS ONE*, *8*(6), e65017.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

Fan, J., & Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association*, *99*(467), 710–723.

Faraway, J. J. (2016). *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models*. CRC Press.

Fenske, N., Fahrmeir, L., Hothorn, T., Rzehak, P., & Höhle, M. (2013). Boosting structured additive quantile regression for longitudinal childhood obesity data. *The International Journal of Biostatistics*, *9*(1), 1–18.

Fenske, N., Kneib, T., & Hothorn, T. (2011). Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. *Journal of the American Statistical Association*, *106*(494), 494–510.

Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (2008). *Longitudinal data analysis*. CRC Press.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2012). *Applied longitudinal analysis*, vol. 998. John Wiley & Sons.

Fornaroli, R., Cabrini, R., Sartori, L., Marazzi, F., Vracevic, D., Mezzanotte, V., Annala, M., & Canobbio, S. (2015). Predicting the constraint effect of environmental characteristics on macroinvertebrate density and diversity using quantile regression mixed model. *Hydrobiologia*, *742*(1), 153–167.

Fox, J. (2000). *Nonparametric simple regression: Smoothing scatterplots*. 130. Sage.

Fox, J., & Monette, G. (2002). *An R and S-Plus companion to applied regression*. Sage.

Friedman, J. H., & Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, *76*(376), 817–823.

Fu, L., & Wang, Y.-G. (2012). Quantile regression for longitudinal data with a working correlation model. *Computational Statistics & Data Analysis*, *56*(8), 2526–2538.

Galarza, C. (2015). *Quantile regression for mixed-effects models*. Ph.D. thesis.

Galarza, C. E., Castro, L. M., Louzada, F., & Lachos, V. H. (2020). Quantile regression for nonlinear mixed effects models: A likelihood based perspective. *Statistical Papers*, *61*(3), 1281–1307.

Galarza, C. E., & Galarza, M. C. E. (2015). Package 'ald'. *Communications in Statistics-Theory and Methods*, *34*(9-10), 1867–1879.

Galarza, C. E., Lachos, V. H., & Bandyopadhyay, D. (2017). Quantile regression in linear mixed models: A stochastic approximation EM approach. *Statistics and its Interface*, *10*(3), 471.

Galarza, C. E., et al. (2015). Quantile regression for mixed-effects models= Regressão quantílica para modelos de efeitos mistos.

Galecki, A. T. (1994). General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics-Theory and Methods*, *23*(11), 3105–3119.

Galvao Jr, A. F. (2011). Quantile regression for dynamic panel data with fixed effects. *Journal of Econometrics*, *164*(1), 142–157.

Garrett, N., Norman, E., Leask, K., Naicker, N., Asari, V., Majola, N., Karim, Q. A., & Karim, S. S. A. (2018). Acceptability of early antiretroviral therapy among South African women. *AIDS and Behavior*, *22*(3), 1018–1024.

Gbur, E. E., Stroup, W. W., McCarter, K. S., Durham, S., Young, L. J., Christman, M., West, M., & Kramer, M. (2012). *Generalized linear mixed models*. American Society of Agronomy, Crop Science Society of America, Soil Science.

Gelfand, A. E. (2000). Gibbs sampling. *Journal of the American Statistical Association*, *95*(452), 1300–1304.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., & Smith, A. F. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association*, *85*(412), 972–985.

Gelfand, A. E., & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*(410), 398–409.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6), 721–741.

Geneva (2017). Ending AIDS: Progress towards the 90-90-90 targets. *Joint United Nations Programme on HIV: Geneva, Switzerland*.

Gentle, J. E. (2009). *Computational Statistics*. Springer.

Geraci, M. (2019). Additive quantile regression for clustered data with an application to children's physical activity. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *68*(4), 1071–1089.

Geraci, M., & Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics*, *8*(1), 140–154.

Geraci, M., & Bottai, M. (2014). Linear quantile mixed models. *Statistics and Computing*, *24*(3), 461–479.

Geraci, M., et al. (2014). Linear quantile mixed models: The lqmm package for laplace quantile regression. *Journal of Statistical Software*, *57*(13), 1–29.

Gilchrist, W. (2000). *Statistical modelling with quantile functions*. CRC Press.

Gill, J., & Torres, M. (2019). *Generalized linear models: A unified approach*, vol. 134. Sage Publications, Incorporated.

Girma, S., & Görg, H. (2005). Foreign direct investment, spillovers and absorptive capacity: Evidence from quantile regressions.

Golub, G. H., & Welsch, J. H. (1969). Calculation of Gauss quadrature rules. *Mathematics of Computation*, *23*(106), 221–230.

Green, P. J., & Silverman, B. W. (1993). *Nonparametric regression and generalized linear models: A roughness penalty approach*. CRC Press.

Grégoire, T. G., Schabenberger, O., & Barrett, J. P. (1995). Linear modelling of irregularly spaced, unbalanced, longitudinal data from permanent-plot measurements. *Canadian Journal of Forest Research*, *25*(1), 137–156.

Guide, S. U. (2008). SAS/ETS 9.2 User's Guide.

Gujarati, D. (2014). *Econometrics by example*. Palgrave Macmillan.

Hahn, J. (1995). Bootstrapping quantile regression estimators. *Econometric Theory*, (pp. 105–121).

Hannan, E. J., & Quinn, B. G. (1979). The Determination of the Order of an Autoregression. *Journal of the Royal Statistical Society: Series B (Methodological)*, *41*(2), 190–195.

Hardin, J., & Hilbe, J. (2003). Generalized Estimating Equations. *Chapman & Hall/CRC: Boca Raton, Florida*.

Harezlak, J., Ruppert, D., & Wand, M. P. (2018). *Semiparametric regression with R*. Springer.

Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C. E., Robinson, B. S., Hodgson, D. J., & Inger, R. (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, *6*, e4794.

Harville, D., & Callanan, T. (1990). Computational aspects of likelihood-based inference for variance components. In *Advances in statistical methods for genetic improvement of livestock*, (pp. 136–176). Springer.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, *72*(358), 320–338.

Hastie, T., & Tibshirani, R. (1990). Generalized additive models. CRC Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hastie, T. J. (2017). *Generalized additive models*. Routledge.

He, X., Ng, P., & Portnoy, S. (1998). Bivariate quantile smoothing splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *60*(3), 537–550.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*, vol. 451. John Wiley & Sons.

Hendricks, W., & Koenker, R. (1992). Hierarchical spline models for conditional quantiles and the demand for electricity. *Journal of the American statistical Association*, *87*(417), 58–68.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

Hilbe, J. M. (2014). *Modeling count data*. Cambridge University Press.

Hofer, A. (1998). Variance component estimation in animal breeding: A review. *Journal of Animal Breeding and Genetics*, *115*(1-6), 247–265.

Hong, H. G., Christiani, D. C., & Li, Y. (2019). Quantile regression for survival data in modern cancer research: Expanding statistical tools for precision medicine. *Precision Clinical Medicine*, *2*(2), 90–99.

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2010). *Multilevel analysis: Techniques and Applications*. Routledge.

Huang, Q., Zhang, H., Chen, J., & He, M. (2017). Quantile regression models and their applications:A review. *J Biom Biostat*, *8*(10.4172), 2155–6180.

Hughson, G. (2017). The Foundation for AIDS Research. 'Statistics: Women and HIV/AIDS'. Accessed: 2020-7-12.
URL https://www.aidsmap.com/about-hiv/cd4-cell-counts

Jank, W. (2006). Implementing and diagnosing the stochastic approximation EM algorithm. *Journal of Computational and Graphical Statistics*, *15*(4), 803–829.

Jiang, J. (2007). *Linear and generalized linear mixed models and their applications*. Springer Science & Business Media.

Jones, R. H. (1993). *Longitudinal data with serial correlation: A state-space approach*. CRC Press.

Karim, S., Williamson, C., & Garrett, N. (2017). Viral set point and clinical disease progression: The role of immunological, genetic and viral factors over the course of disease and during antiretroviral therapy. CAP002: Acute Infection Study. Accessed: 2019-09-12.
URL https://www.caprisa.org/Pages/CAPRISAStudies

Kassutto, S., & Rosenberg, E. S. (2004). Primary HIV type 1 Infection. *Clinical Infectious Diseases*, *38*(10), 1447–1453.

Kincaid, C. (2005). Guidelines for selecting the covariance structure in mixed model analysis. In *Proceedings of the thirtieth annual SAS users group international conference*, vol. 30, (pp. 198–130). SAS Institute Inc Cary NC.

Kleinbaum, D. G., Klein, M., et al. (2012). *Survival analysis: A self-learning text*, vol. 3. Springer.

Kleiner, A., Talwalkar, A., Sarkar, P., & Jordan, M. I. (2014). A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, (pp. 795–816).

Knight, C. A., & Ackerly, D. D. (2002). Variation in nuclear DNA content across environmental gradients: A quantile regression analysis. *Ecology Letters*, *5*(1), 66–76.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, *91*(1), 74–89.

Koenker, R. (2005a). *Quantile regression*. Cambridge University Press.

Koenker, R. (2005b). Quantile regression: Econometric society monographs. *Cambridge University Press; Cambridge and New York*, (38).

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, (pp. 33–50).

Koenker, R., & Geling, O. (2001). Reappraising medfly longevity: A quantile regression survival analysis. *Journal of the American Statistical Association*, *96*(454), 458–468.

Koenker, R., & Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives*, *15*(4), 143–156.

Koenker, R., & Machado, J. A. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, *94*(448), 1296–1310.

Koenker, R., Ng, P., & Portnoy, S. (1994). Quantile smoothing splines. *Biometrika*, *81*(4), 673–680.

Koenker, R., Portnoy, S., & Ng, P. (1992). Nonparametric estimation of conditional quantile functions. *Dodge, Y.(Ed)*.

Koenker, R., et al. (2008). Censored quantile regression redux. *Journal of Statistical Software*, *27*(6), 1–25.

Koenker, R., et al. (2011). Additive models for quantile regression: Model selection and confidence bandaids. *Brazilian Journal of Probability and Statistics*, *25*(3), 239–262.

Kohn, R., Ansley, C. F., & Tharm, D. (1991). The performance of cross-validation and maximum likelihood estimators of spline smoothing parameters. *Journal of the American Statistical Association*, *86*(416), 1042–1050.

Kotz, S., Kozubowski, T., & Podgorski, K. (2012). *The Laplace distribution and generalizations: A revisit with applications to Communications, Economics, Engineering, and Finance*. Springer Science & Business Media.

Kotz, S., Kozubowski, T. J., & Podgórski, K. (2002). Maximum likelihood estimation of asymmetric Laplace parameters. *Annals of the Institute of Statistical Mathematics*, *54*(4), 816–826.

Kowalchuk, R. K., Keselman, H., Algina, J., & Wolfinger, R. D. (2004). The analysis of repeated measurements with mixed-model adjusted F tests. *Educational and Psychological Measurement*, *64*(2), 224–242.

Kozubowski, T. J., & Nadarajah, S. (2010). Multitude of Laplace distributions. *Statistical Papers*, *51*(1), 127.

Kozumi, H., & Kobayashi, G. (2011). Gibbs sampling methods for Bayesian quantile regression. *Journal of Statistical Computation and Simulation*, *81*(11), 1565–1578.

Kuan, C.-M. (2007). An introduction to quantile regression. *Institute of Economics Academica Sinica*.

Kuhn, E., & Lavielle, M. (2004). Coupling a Stochastic Approximation version of EM with an MCMC procedure. *ESAIM: Probability and Statistics*, *8*, 115–131.

Kuhn, E., & Lavielle, M. (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics & Data Analysis*, *49*(4), 1020–1038.

Lachos, V. H., Chen, M.-H., Abanto-Valle, C. A., & Azevedo, C. L. (2015). Quantile regression for censored mixed-effects models with applications to HIV studies. *Statistics and its Interface*, *8*(2), 203.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*(1), 1–14.

Lawless, J. F. (1987). Negative binomial and mixed poisson regression. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, (pp. 209–225).

Li, X., Ding, H., Geng, W., Liu, J., Jiang, Y., Xu, J., Zhang, Z., & Shang, H. (2019). Predictive effects of body mass index on immune reconstitution among HIV-Infected HAART users in China. *BMC Infectious Diseases*, *19*(1), 373.

Li, Y., Liu, Y., & Zhu, J. (2007). Quantile regression in reproducing kernel hilbert spaces. *Journal of the American Statistical Association*, *102*(477), 255–268.

Liang, H., Wu, H., & Carroll, R. J. (2003). The relationship between virologic and immunologic responses in AIDS clinical research using mixed-effects varying-coefficient models with measurement error. *Biostatistics*, *4*(2), 297–312.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22.

Lin, C.-Y., Bondell, H., Zhang, H. H., & Zou, H. (2013). Variable selection for non-parametric quantile regression via smoothing spline analysis of variance. *Stat*, *2*(1), 255–268.

Lin, X., & Zhang, D. (1999). Inference in generalized additive mixed modelsby using smoothing splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(2), 381–400.

Lindsey, J. K. (2001). *Nonlinear models in medical statistics*. Oxford University Press on Demand.

Lipsitz, S. R., Fitzmaurice, G. M., Molenberghs, G., & Zhao, L. P. (1997). Quantile regression methods for longitudinal data with drop-outs: Application to CD4 cell counts of patients infected with the Human Immunodeficiency Virus. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *46*(4), 463–476.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Oliver, S. (2006). *SAS for mixed models*. SAS Publishing.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data*, vol. 793. John Wiley & Sons.

Liu, W., & Cela, J. (2008). Count data models in SAS. In *SAS Global Forum*, vol. 317, (pp. 1–12).

Liu, X. (2015). *Methods and applications of longitudinal data analysis*. Elsevier.

Liu, Y., & Bottai, M. (2009). Mixed-effects models for conditional quantiles with longitudinal data. *The International Journal of Biostatistics*, *5*(1).

Lord, D., Park, B.-J., & Model, P.-G. (2012). Negative binomial regression models and estimation methods. *Probability Density and Likelihood Functions. Texas A&M University, Korea Transport Institute*, (pp. 1–15).

Loy, A., Hofmann, H., & Cook, D. (2017). Model choice and diagnostics for linear mixed-effects models using statistics on street corners. *Journal of Computational and Graphical Statistics*, *26*(3), 478–492.

Machado, J. A. F., & Silva, J. S. (2005). Quantiles for counts. *Journal of the American Statistical Association*, *100*(472), 1226–1237.

Mahmud, M., Abrahamowicz, M., Leffondré, K., & Chaubey, Y. P. (2006). Selecting the optimal transformation of a continuous covariate in Cox's regression: Implications for hypothesis testing. *Communications in Statistics-Simulation and Computation*, *35*(1), 27–45.

Mamouridis, V. (2011). Additive mixed models applied to the study of red shrimp landings: Comparison between frequentist and bayesian perspectives. *Universidad de Coruña. Departamento de Matemáticas. España*.

Manning, W. G., et al. (1998). The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of Health Economics*, *17*(3), 283–295.

Marino, M. F., & Farcomeni, A. (2015). Linear quantile regression models for longitudinal experiments: An overview. *Metron*, *73*(2), 229–247.

Matousek, J., & Gärtner, B. (2007). *Understanding and using Linear Programming*. Springer Science & Business Media.

McArdle, B. H., & Anderson, M. J. (2004). Variance heterogeneity, transformations, and models of species abundance: A cautionary tale. *Canadian Journal of Fisheries and Aquatic Sciences*, *61*(7), 1294–1302.

McCullagh, P., & Nelder, J. A. (1989). Generalized Linear Models 2nd Edition Chapman and Hall. *London, UK*.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, *92*(437), 162–170.

McCulloch, C. E., & Neuhaus, J. M. (2014). Generalized linear mixed models. *Wiley StatsRef: Statistics Reference Online*.

McCulloch, C. E., Searle, S. R., & Neuhaus, J. M. (2008). Generalized, linear, and mixed models. hoboken. *NJ: Wiley. QA*, *279*, M38.

McLachlan, G. J., & Krishnan, T. (2007). *The EM algorithm and extensions*, vol. 382. John Wiley & Sons.

Melesse, S. F., & Zewotir, T. (2017). Modelling the effect of tree age and climatic factors on the stem radial growth of juvenile eucalypt clones. *Bulletin of the Transilvania University of Brasov. Forestry, Wood Industry, Agricultural Food Engineering. Series II*, *10*(1).

Melesse, S. F., & Zewotir, T. (2020). Additive mixed models to study the effect of tree age and climatic factors on stem radial growth of Eucalyptus trees. *Journal of Forestry Research*, *31*(2), 463–473.

Menard, S. (2002). *Applied logistic regression analysis*, vol. 106. Sage.

Meza, C., Osorio, F., & De la Cruz, R. (2012). Estimation in nonlinear mixed-effects models using heavy-tailed distributions. *Statistics and Computing*, *22*(1), 121–139.

Miranda, A. (2007). QCOUNT: Stata program to fit quantile regression models for count data.

Miranda, A. (2008). Planned fertility and family background: A quantile regression for counts analysis. *Journal of Population Economics*, *21*(1), 67–81.

Mirnezami, R., Nicholson, J., & Darzi, A. (2012). Preparing for precision medicine. *N Engl J Med*, *366*(6), 489–491.

Mlisana, K., Werner, L., Garrett, N. J., McKinnon, L. R., van Loggerenberg, F., Passmore, J.-A. S., Gray, C. M., Morris, L., Williamson, C., & Abdool Karim, S. S. (2014). Rapid disease progression in HIV-1 subtype C-Infected South African Women. *Clinical Infectious Diseases*, *59*(9), 1322–1331.

Molenberghs, G., & Kenward, M. (2007). *Missing data in clinical studies*, vol. 61. John Wiley & Sons.

Molenberghs, G., & Verbeke, G. (2000). *Linear mixed models for longitudinal data*. Springer.

Molenberghs, G., & Verbeke, G. (2006). *Models for discrete longitudinal data*. Springer Science & Business Media.

Molenberghs, G., Verbeke, G., & Demétrio, C. G. (2007). An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime Data Analysis*, *13*(4), 513–531.

Molenberghs, G., Verbeke, G., Demétrio, C. G., & Vieira, A. M. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science*, *25*(3), 325–347.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

Moosa, Y., Tanko, R. F., Ramsuran, V., Singh, R., Madzivhandila, M., Yende-Zuma, N., Abrahams, M.-R., Selhorst, P., Gounder, K., Moore, P. L., et al. (2018). Case report: Mechanisms of HIV elite control in two African women. *BMC Infectious Diseases*, *18*(1), 1–7.

Morel, J. G., & Neerchal, N. (2012). *Overdispersion models in SAS*. SAS Institute.

Muir, P. R., Wallace, C. C., Done, T., & Aguirre, J. D. (2015). Limited scope for latitudinal extension of reef corals. *Science*, *348*(6239), 1135–1138.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, *33*(3), 341–365.

Müller, H.-G. (2012). *Nonparametric regression analysis of longitudinal data*, vol. 46. Springer Science & Business Media.

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, *135*(3), 370–384.

Nieuwenhuis, R., Te Grotenhuis, H., & Pelzer, B. (2012). Influence.ME: Tools for detecting influential data in mixed effects models.

Noufaily, A., & Jones, M. (2013). Parametric quantile regression based on the generalized gamma distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *62*(5), 723–740.

Palermo, B., Bosch, R. J., Bennett, K., & Jacobson, J. M. (2011). Body mass index and CD4+ T-lymphocyte recovery in HIV-Infected men with viral suppression on antiretroviral therapy. *HIV Clinical Trials*, *12*(4), 222–227.

Patel, D. E., Geraci, M., & Cortina-Borja, M. (2016). Modeling normative kinetic perimetry isopters using mixed-effects quantile regression. *Journal of Vision*, *16*(6), 7–7.

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, *58*(3), 545–554.

Peterson, M. D., & Krishnan, C. (2015). Growth charts for muscular strength capacity with quantile regression. *American Journal of Preventive Medicine*, *49*(6), 935–938.

Pinheiro, J., & Bates, D. (2006). *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media.

Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, *4*(1), 12–35.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., Solenberger, P., et al. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, *27*(1), 85–96.

Rawlings, J. O., Pantula, S. G., & Dickey, D. A. (2001). *Applied regression analysis: A research tool*. Springer Science & Business Media.

Reich, B. J., Bondell, H. D., & Wang, H. J. (2010). Flexible bayesian quantile regression for independent and clustered data. *Biostatistics*, *11*(2), 337–352.

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *54*(3), 507–554.

Robinson, G. K., et al. (1991). That blup is a good thing: The estimation of random effects. *Statistical Science*, *6*(1), 15–32.

Rosenberg, E. S., Altfeld, M., Poon, S. H., Phillips, M. N., Wilkes, B. M., Eldridge, R. L., Robbins, G. K., Richard, T., Goulder, P. J., & Walker, B. D. (2000). Immune control of HIV-1 after early treatment of acute infection. *Nature*, *407*(6803), 523–526.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*. 12. Cambridge University Press.

SAS, I. (2014). Sas/stat r 13.2 users guide. *Cary, North Carolina: SAS Institute Inc*.

Schabenberger, O. (2005). Mixed model influence diagnostics. In *SUGI*, vol. 29, (pp. 189–29). Citeseer.

Schabenberger, O., & Gotway, C. A. (2017). *Statistical methods for spatial data analysis*. CRC Press.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC Press.

Schwarz, G., et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Searle, S. (1997). Linear models. Wiley Classics Library.

Searle, S. R. (1982). *Matrix algebra useful for statistics*.

Searle, S. R., Casella, G., & McCulloch, C. E. (2009). *Variance components*, vol. 391. John Wiley & Sons.

Searle, S. R., & Gruber, M. H. (2016). *Linear models*. John Wiley & Sons.

Shadish, W. R., Zuur, A. F., & Sullivan, K. J. (2014). Using generalized additive (mixed) models to analyze single case designs. *Journal of School Psychology*, *52*(2), 149–178.

Sherwood, B., Wang, L., & Zhou, X.-H. (2013). Weighted quantile regression for analyzing health care cost data with missing covariates. *Statistics in Medicine*, *32*(28), 4967–4979.

Shisana, O., Rehle, T., Simbayi, L. C., Zuma, K., Jooste, S., Zungu, N., Labadarios, D., & Onoya, D. (2014). South African National HIV Prevalence, Incidence and Behaviour Survey, 2012.

Shoukri, M., Asyali, M., VanDorp, R., & Kelton, D. (2004). The Poisson inverse Gaussian regression model in the analysis of clustered counts data. *Journal of Data Science*, *2*(1), 17–32.

Shoukri, M. M. (2018). *Analysis of Correlated Data with SAS and R*. CRC Press.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society: Series B (Methodological)*, *47*(1), 1–21.

Song, X., Li, G., Zhou, Z., Wang, X., Ionita-Laza, I., & Wei, Y. (2017). QRank: A novel quantile regression tool for eQTL discovery. *Bioinformatics*, *33*(14), 2123–2130.

Stone, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, (pp. 689–705).

Stroup, W. W. (2012). *Generalized linear mixed models: Modern concepts, methods and applications*. CRC Press.

Stroup, W. W. (2015). Rethinking the analysis of non-normal data in plant and soil science. *Agronomy Journal*, *107*(2), 811–827.

Stukel, T. A. (1993). Comparison of methods for the analysis of longitudinal interval count data. *Statistics in Medicine*, *12*(14), 1339–1351.

Tao, H., Palta, M., Yandell, B. S., & Newton, M. A. (1999). An estimation method for the semiparametric mixed effects model. *Biometrics*, *55*(1), 102–110.

Taris, T. W. (2000). *A primer in longitudinal data analysis*. Sage.

Thall, P. F. (1988). Mixed Poisson likelihood regression models for longitudinal interval count data. *Biometrics*, (pp. 197–209).

Tian, Y., Wang, L., Tang, M., & Tian, M. (2020). Likelihood-based quantile mixed effects models for longitudinal data with multiple features via MCEM algorithm. *Communications in Statistics-Simulation and Computation*, *49*(2), 317–334.

Twisk, J. W. (2013). *Applied longitudinal data analysis for epidemiology: A practical guide*. Cambridge University Press.

UN (2014). Message from UN Women's Executive Director for World AIDS Day, 1 December 2014. Accessed: 2019-10-07.
URL https://bit.ly/3xp0AM7

UNAIDS (2017). By the Numbers. UNAIDS, Geneva, 2016.

UNAIDS (2019). Global HIV and AIDS Statistics. Accessed: 2020-10-12.
URL https://www.avert.org/global-hiv-and-aids-statistics

Van der Meer, T., Te Grotenhuis, M., & Pelzer, B. (2010). Influential cases in multilevel modeling: A methodological comment. *American Sociological Review*, *75*(1), 173–178.

Van Loggerenberg, F., Mlisana, K., Williamson, C., Auld, S. C., Morris, L., Gray, C. M., Karim, Q. A., Grobler, A., Barnabas, N., Iriogbe, I., et al. (2008). Establishing a cohort at high risk of HIV Infection in South Africa: Challenges and experiences of the CAPRISA 002 Acute Infection Study. *PLOS ONE*, *3*(4), e1954.

Vanderbei, R. J. (2020). *Linear Programming: Foundations and extensions*, vol. 285. Springer Nature.

Verbeke, G., & Molenberghs, G. (2009). Linear mixed models for longitudinal data. *Springer Science & Business Media*.

Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *The Annals of Statistics*, (pp. 1378–1402).

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, *61*(3), 439–447.

Wei, Y., Pere, A., Koenker, R., & He, X. (2006). Quantile regression methods for reference growth charts. *Statistics in Medicine*, *25*(8), 1369–1382.

Weisberg, S. (2005). *Applied linear regression*, vol. 528. John Wiley & Sons.

Weiss, R. E. (2005). *Modeling longitudinal data*. Springer Science & Business Media.

West, B. T., Welch, K. B., & Galecki, A. T. (2014). *Linear mixed models: A practical guide using statistical software*. CRC Press.

Whelan, D. (1999). *Gender and HIV/AIDS: Taking stock of research and programmes*. UNAIDS.

WHO (2010). *AIDS epidemic update: December 2009*. WHO Regional Office Europe.

WHO, UNAIDS, U., et al. (2007). *Women, ageing and health: A framework for action: Focus on gender*. Geneva: World Health Organization.

WHO, UNAIDS, U., et al. (2008). Epidemiological fact sheet on HIV and AIDS; Core data on epidemiology and response, South Africa. *Geneva: WHO*, (pp. 1–19).

Wichitaksorn, N., Choy, S. B., & Gerlach, R. (2014). A generalized class of skew distributions and associated robust quantile regression models. *Canadian Journal of Statistics*, *42*(4), 579–596.

Winkelmann, R. (2006). Reforming health care: Evidence from quantile regressions for counts. *Journal of Health Economics*, *25*(1), 131–145.

Winkelmann, R. (2008). *Econometric analysis of count data*. Springer Science & Business Media.

Wolfinger, R. D. (1996). Heterogeneous variance: Covariance structures for repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics*, (pp. 205–230).

Wood, S., & Wood, M. S. (2015). Package 'mgcv'. *R Package Version*, *1*, 29.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *65*(1), 95–114.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. CRC Press.

Wu, H., & Zhang, J.-T. (2006). *Nonparametric regression methods for longitudinal data analysis: mixed-effects modeling approaches*, vol. 515. John Wiley & Sons.

Xiang, D. (2001). Fitting generalized additive models with the gam procedure. In *SUGI Proceedings*, (pp. 256–26). Citeseer.

Yirga, A., Ayele, D., & Melesse, S. (2018). Application of Quantile Regression: Modeling Body Mass Index in Ethiopia. *The Open Public Health Journal*, *11*, 221–233.
URL https://benthamopen.com/FULLTEXT/TOPHJ-11-221

Yirga, A. A. (2018). *Statistical models to study the BMI of under five children in Ethopia, 10 April 2019*.
URL https://bit.ly/2TApTwp

Yirga, A. A., Melesse, S. F., Mwambi, H. G., & Ayele, D. G. (2020a). Modelling CD4 counts before and after HAART for HIV infected patients in KwaZulu-Natal South Africa. *African Health Sciences*, *20*(4), 1546–61.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8351836/

Yirga, A. A., Melesse, S. F., Mwambi, H. G., & Ayele, D. G. (2020b). Negative binomial mixed models for analyzing longitudinal CD4 count data. *Scientific Reports*, *10*(1), 1–15.
URL https://rdcu.be/b8bnO

Yirga, A. A., Melesse, S. F., Mwambi, H. G., & Ayele, D. G. (2021a). Additive quantile mixed effects modelling with application to longitudinal CD4 count data. *Scientific reports*, *11*(1), 1–12.
URL https://rdcu.be/cxrJ3

Yirga, A. A., Melesse, S. F., Mwambi, H. G., & Ayele, D. G. (2021b). Analyzing longitudinal CD4 count of HIV-infected patients using generalized additive mixed-effects model. Under Review.

Yirga, A. A., Melesse, S. F., Mwambi, H. G., & Ayele, D. G. (2022). Application of quantile mixed-effects model in modeling CD4 count from HIV-infected patients in KwaZulu-Natal South Africa. *BMC Infectious Diseases*, *22*(1), 1–11.
URL https://rdcu.be/cEluO

Yu, K., Lu, Z., & Stander, J. (2003). Quantile regression: Applications and current research areas. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *52*(3), 331–350.

Yu, K., & Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, *54*(4), 437–447.

Yu, K., & Zhang, J. (2005). A three-parameter asymmetric Laplace distribution and its extension. *Communications in Statistics-Theory and Methods*, *34*(9-10), 1867–1879.

Zeger, S. L., & Diggle, P. J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics*, (pp. 689–699).

Zeger, S. L., & Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, (pp. 121–130).

Zewotir, T. (2008). Multiple cases deletion diagnostics for linear mixed models. *Communications in Statistics-Theory and Methods*, *37*(7), 1071–1084.

Zewotir, T., & Galpin, J. S. (2005). Influence diagnostics for linear mixed models. *Journal of Data Science*, *3*(2), 153–177.

Zhang, D., Lin, X., Raz, J., & Sowers, M. (1998). Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association*, *93*(442), 710–719.

Zhang, X., Mallick, H., Tang, Z., Zhang, L., Cui, X., Benson, A. K., & Yi, N. (2017). Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinformatics*, *18*(1), 4.

Zhang, X., Pei, Y.-F., Zhang, L., Guo, B., Pendegraft, A. H., Zhuang, W., & Yi, N. (2018). Negative binomial mixed models for analyzing longitudinal microbiome data. *Frontiers in Microbiology*, *9*, 1683.

Zimmerman, D. L., & Harville, D. A. (1991). A random field approach to the analysis of field-plot experiments and other spatial experiments. *Biometrics*, (pp. 223–239).

Zuur, A., Ieno, E. N., Walker, N., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. Springer Science & Business Media.

Zwillinger, D. (2002). *CRC Standard Mathematical Tables and Formulae*. CRC Press.

Appendix A: Codes

Sample statistical software codes used to analyze the data examples

############################## **Chapter 2** ##############################

Code 7.1: SAS and R codes for the data example in Section 2.6

```
libname mylib "C:\Users\216065934\Desktop\PhD_December";

PROC IMPORT OUT= WORK.dataMFile2
DATAFILE= "C:\Users\216065934\Desktop\PhD_December\Modified-
   Quantile"
DBMS=SPSS REPLACE;
RUN;

/***********Exploratory Data Analysis************/

/*Overview of Data*/
proc contents data=WORK.acute;
run;

/*Freq, Means, Univariate*/

proc freq data=WORK.acute;
run;

proc univariate data=WORK.acute;
run;

proc means data=WORK.acute min mean max;
var p27v12 viralload;
run;
/************Testing normality*******************/
ods html;
ods graphics on;
```

```
*title c=bib height=1 'Testing normality of the actual response
   variable (CD4 count) using histogram';
proc sgplot data=dataM;
histogram CD4_count;
density CD4_count / type=normal;
run;
ods graphics off;
ods html close;

ods html;
ods graphics on;
*title c=bib height=1 'Testing normality of the square root
   transformed CD4 count using histogram';
proc sgplot data=dataM;
histogram CD4_count;
density CD4_count / type=normal;
run;
ods graphics off;
ods html close;

/**********Calculating time variable from the data set*****/

data DataPostTime;
set time;
by p1v7;
retain date_lag;
if first.p1v7 then do;
time = 0;
date_lag = p27v8;
end;
else do;
time = (p27v8 - date_lag);
date_lag = p27v8;
end;
run;

proc export data=DataPostTime outfile="C:\Users\216065934\Desktop\
   PhD\working data\DataPostTime.sav"
dbms=spss replace;
run;
data DataTotalTime;
set DataPostTime;
by p1v7;
retain total_lag;
```

```
if first.p1v7 then do;
TotalTime= 0;
total_lag = time;
end;
else do;
TotalTime = (time + total_lag);
total_lag = (time + total_lag);
end;
run;
data DataTotalTimeWeek;
set DataTotalTime;
TotalTimeWeek = TotalTime/7;
run;

data DataTotalTimeMonth;
set DataTotalTimeWeek;
TotalTimeMonth = TotalTime/30;
run;

data DataTotalTimeYear;
set DataTotalTimeMonth;
TotalTimeYear = TotalTime/365;
run;

proc export data=DataTotalTimeYear outfile="C:\Users\216065934\
   Desktop\PhD\working data\CAP002DataWithTime.sav"
dbms=spss replace;
run;


/************************Mean response profile plot************/
ods html;
ods graphics on;
proc sgplot data=dataM;
vline TotalTimeYear / response=p27v12 stat=mean group=ART
limitstat=stderr;
run;
ods graphics on;
ods html close;

ods html;
ods graphics on;
proc sgplot data=dataM;
vline TotalTimeYear / response=SQRT_CD4 stat=mean group=ART
```

```
limitstat=stderr;
run;
ods graphics on;
ods html close;

/***************Baseline data**********/

data acuteCAP002baseline;
set acuteCAP002;
if dfseq=2000; *where "2000" is a vistcode;
run;
proc sort data=acuteCAP002baseline nodupkey;
by p1v7;
run;

proc export data=acuteCAP002baseline outfile="C:\Users\216065934\
   Desktop\PhD\Ddata\acuteCAP002baseline.sav"
dbms=spss replace;
run;
/*************** Selecting random samples************************/

title 'Simple Random Sampling';
proc surveyselect data=Baseline
method=srs n=15 out=AcuteSRS1;
run;

/***Individual profiles plot for CD4+ count in 15 randomly selected
    individuals by Pre and Post ART initiation group***/

ods html;
ods graphics on;
proc sgpanel data=dataMS;
panelby ART / spacing=10;
series y=p27v12 x=TotalTimeMonth /group=ParticipantID;
run;
ods graphics on;
ods html close;

/*************Figure: 2.3 using R*********/

women1<- read.spss("DATACAPSrS.sav", to.data.frame = TRUE, use.
   missings = TRUE)

head(women1)
```

```
str(women1)

pp <- ggplot(data = women1, aes(x = TotalTimeMonth, y = SQRT_CD4 ,
    group = ParticipantID))

pp + geom_line() + facet_wrap(~ ParticipantID)
ppp + geom_line() + stat_summary(aes(group = 1), geom = "point",
    fun.y = mean, shape = 17, size = 3) + facet_wrap( ~
    ParticipantID)

/***Selecting best random effect model and covariance structure***/

ods html;
ods graphics on;
proc mixed data=dataM1 covtest asycov asycorr ic method=REML  PLOTS
    (MAXPOINTS=none)= all;
class PID ART;
model SQRT_CD4= TotalTimeMonth SQRT_Time ART /s cl ddfm=bw;
random int TotalTimeMonth SQRT_Time/subject=PID g gcorr v vcorr
    solution type=un; *type=un; *type=cs; *type=AR(1); *type=TOEP;
run;
ods graphics off;
ods html close;


/***Model 1: Intercept, time and SQRT_Time (Random intercept and
    Random slopes model)***/

ods pdf file="C:\Users\216065934\Desktop\MixedModeldocument\
    ProcMixedOutput.pdf";
ods html;
ods graphics on;
proc mixed data=dataM1 covtest asycov asycorr ic  method=REML;
class   PID BMI_category(ref="Normal weight") ART
    Baseline_VL_category(ref="Low VL") NumberOfSexPartner Agegroup(
    ref="<20") Educational_status Residence;
model SQRT_CD4= TotalTimeMonth SQRT_Time  BMI_category ART
    Baseline_VL_category  NumberOfSexPartner Agegroup
    Educational_status Residence/s cl ddfm=bw outp=predmixed;
random int TotalTimeMonth SQRT_Time/subject=PID solution type=un g
    gcorr v vcorr; *type=un; *type=cs; *type=AR(1); *type=TOEP;
run;
ods graphics off;
ods html close;
```

```
ods pdf close;
ods pdf close;

/***To see R-square and the fitted equation (Heat map)***/
ods html;
ods graphics on;
goptions reset=global gunit=pct ftext=swissb htitle=5 htext=3 ;
goptions rotate=landscape gsfname=graph2 gsfmode=append ;
title 'Fit Plot for Predicted CD4+ count';
proc reg data=Predicted2 PLOTS(MAXPOINTS=none);
model Fitted_average_CD4_count = Observed_CD4_count;
plot Fitted_average_CD4_count*Observed_CD4_count/conf pred
modelfont=swiss modellab='Sample plot' modelht=4
statfont=swiss statht=4;
run;
ods graphics off;
ods html close;

/***************Diagnostics and influence analysis**************/
ods html;
ods graphics on;
proc mixed data=dataM1 method=REML PLOTS(MAXPOINTS=none)=all;
class   PID BMI_category(ref="Normal weight") ART
   Baseline_VL_category2(ref="Low VL") NumberOfSexPartner Agegroup(
   ref="<20") Educational_status Residence_mod;
model SQRT_CD4= TotalTimeMonth SQRT_Time  BMI_category ART
   Baseline_VL_category2  NumberOfSexPartner Agegroup
   Educational_status Residence_mod/s cl ddfm=bw influence(effect=
   PID iter=5 est);
random int TotalTimeMonth SQRT_Time/subject=PID solution type=un;
run;
ods graphics off;
ods html close;



/***Diagnostics and influence analysis with spatial covariance
   structure***/

ods html;
ods graphics on;
proc mixed data=dataM1 method=REML PLOTS(MAXPOINTS=none)=all;
class   PID TimeMonth_factor BMI_category(ref="Normal weight") ART
```

```
    Baseline_VL_category2(ref="Low VL") NumberOfSexPartner Agegroup(
    ref="<20") Educational_status Residence_mod;
model SQRT_CD4= TotalTimeMonth SQRT_Time  BMI_category ART
    Baseline_VL_category2  NumberOfSexPartner Agegroup
    Educational_status Residence_mod/s cl ddfm=bw influence(effect=
    PID iter=5 est);
repeated TimeMonth_factor/type=sp(exp)(TotalTimeMonth) local sub=
    PID; *sp(exp), sp(sph), sp(gau);
*random int TotalTimeMonth SQRT_Time/subject=PID solution type=un;
run;
ods graphics off;
ods html close;
```

############################## **Chapter 3** ##############################

Code 7.2: SAS code for the data example in Section 3.7

```
/******************Individual profile plot*******************/

proc sort data=dataM3 out=studyear; *dataM3 is a data set with 17
    randomly selected individuals;
by p27v8 desending p27v8;
format p27v8 year4.;
run;

ods html;
ods graphics on;
proc sgplot data=studyear;
title c=bib height=1 'Individual Profile Plot for CD4+ Count';
xaxis label = "Year";
yaxis label = "Number of CD4+ cell";
series y=p27v12 x=TotalTimeYear/ group=PID;
run;
ods graphics on;
ods html close;

/*******GLMM with Poisson and Negative binomial ditribution******/
ods html;
ods graphics on;
proc glimmix data=MDA4 method=laplace plot=residualpanel(unpack
    ilink)
plot=residualpanel(unpack noilink) plot=studentpanel(unpack noilink
    );
class   PID BMI_category(ref="Normal weight") ART
```

```
    Baseline_VL_category2(ref="Low VL") NumberOfSexPartner Agegroup(
    ref="<20") Educational_status site;
model p27v12= TotalTimeMonth SQRT_Time  BMI_category ART
    Baseline_VL_category2  NumberOfSexPartner Agegroup
    Educational_status site/dist= poisson solution  link=log cl;
random int TotalTimeMonth SQRT_Time/subject=PID  solution type=un;
run;
ods graphics off;
ods html close;

ods html;
ods graphics on;
proc glimmix data=MDA4 method=laplace plot=residualpanel(unpack
    ilink)
plot=residualpanel(unpack noilink) plot=studentpanel(unpack noilink
    );
class   ParticipantID Timemonth_factor Baseline_BMI_category(ref="
    Normal weight") ART Baseline_VL_category2(ref="Low VL")
    Marital_status Agegroup(ref="<20") Educational_attainment
    Residence_mod;
model p27v12= TotalTimeMonth SQRT_of_Time_month
    Baseline_BMI_category ART Baseline_VL_category2  Marital_status
    Agegroup Educational_attainment Residence_mod/dist= negbin
    solution  link=log cl;
random int TotalTimeMonth SQRT_of_Time_month/subject=ParticipantID
    solution type=un;
output out = MDA5 pred=prob ;
run;
ods graphics off;
ods html close;

/***Predicted CD4 profile plot for selected individual using
    Negative binomial mixed model***/

ods html;
ods graphics on;
*title c=bib height=1 'Distribution of Month post infection';
proc sgplot data=MDA5SRS;
title c=bib height=1 'Prediction of Individual Profiles for CD4+
    cell count ';
xaxis label = "Time in Years";
yaxis label = "Predicted CD4 Count";
yaxis min=0 max=1200;
xaxis min=0 max=4;
```

```
where PID between 141 and 205;
reg y=Predicted_CD4_Count x=TotalTimeYear/ group=PID nomarkers;
run;
ods graphics off;
ods html close;

/***Proc MIAnalyse for GLMM with Negative binomial distribution***/

ods html;
ods graphics on;
Proc MI data=MDA4 seed=69301 nimpute=10 out=MA ; *pctmissing(min=5
   max=20);
class  Baseline_BMIndex ART Baseline_VLoad Marital_status Agegroup
   Educational_level Residence_mod;*/ref=first;
fcs  nbiter=10   discrim(Baseline_BMIndex/details) discrim(ART/
   details) discrim(Baseline_VLoad/details) discrim(Marital_status/
   details) discrim(Agegroup/details)
discrim(Educational_level/details) discrim(Residence_mod/details)
   regpmm(p27v12/details) regpmm(sqrtCD4/details) regpmm(
   ViralLoad_mod/details) regpmm(weight/details)
regpmm(height/details) regpmm(age_at_specimen_collection/details)
   regpmm(BMI/details);*regpmm Specifies the predictive mean
   matching method;
var p27v12 sqrtCD4 ViralLoad_mod Baseline_BMIndex ART
   Baseline_VLoad Marital_status Agegroup Educational_level weight
   height age_at_specimen_collection Residence_mod  BMI; *FCS (
   Fully conditional specification);
run; *FCS REGPMM selects the FCS Predicted Mean Matching method to
   impute missing data;
ods graphics off;
ods html close;

ods pdf file="C:\Users\216065934\Desktop_October\GLMM\GLMM_MIFinal.
   pdf";
ods html;
ods graphics on;
proc glimmix data=MA method=laplace;
class    PID Baseline_BMIndex(ref="Normal weight") ART
   Baseline_VLoad(ref="Low VL") Marital_status Agegroup(ref="<20")
   Educational_level Residence_mod;*/ref=first;
model p27v12= TotalTimeMonth SQRT_of_Time_month   Baseline_BMIndex
   ART Baseline_VLoad  Marital_status Agegroup Educational_level
   Residence_mod/dist= negbin  solution  link=log cl covb ddfm=
   residual ;
```

```
random int TotalTimeMonth SQRT_of_Time_month/subject=PID solution
   type=un;
by _imputation_;
ods output parameterestimates=NBparms;
run;
ods graphics off;
ods html close;
ods pdf close;

ods pdf file="C:\Users\216065934\Desktop\PhD_October\GLMM\
   GLMM_MIAnalyzeFinal.pdf";
ods html;
ods graphics on;
proc mianalyze parms=NBparms ; * (effectvar=rowcol); *covb=glmcovb
   edf=218; *(classvar=full) (effectvar=rowcol);
class Baseline_BMIndex ART Baseline_VLoad Marital_status Agegroup
   Educational_level Residence_mod;*/ref=first;
modeleffects intercept TotalTimeMonth SQRT_of_Time_month
   Baseline_BMIndex ART Baseline_VLoad Marital_status Agegroup
   Educational_level Residence_mod;
run;
ods graphics off;
ods html close;
ods pdf close;

/***Examining estimated random effects for the fitted NBMM***/

PROC IMPORT OUT= WORK.RE
DATAFILE= "C:\Users\Student\Desktop\QR\Eamining random effects_NBMM
    - Copy.sav"
DBMS=SPSS REPLACE;
RUN;

ods html;
ods graphics on;
proc univariate data=WORK.RE;
qqplot Intercept Time_Month SQRTof_Time_month;
run;
ods graphics off;
ods html close;

ods html;
ods graphics on;
TITLE 'Histogram of the random effects';
```

```
PROC UNIVARIATE DATA = WORK.RE NOPRINT;
HISTOGRAM Intercept / NORMAL;
RUN;
ods graphics off;
ods html close;

ods html;
ods graphics on;
TITLE 'Histogram of the random effects';
PROC UNIVARIATE DATA = WORK.RE NOPRINT;
HISTOGRAM Time_Month / NORMAL;
RUN;
ods graphics off;
ods html close;

ods html;
ods graphics on;
TITLE 'Histogram of the random effects';
PROC UNIVARIATE DATA = WORK.RE NOPRINT;
HISTOGRAM SQRTof_Time_month / NORMAL;
RUN;
ods graphics off;
ods html close;
```

################################ **Chapter 4** ################################

Code 7.3: R code for the data example in Section 4.4

```
### Densities of an Asymmetric Laplace Distribution using R ####

install.packages("ald")
library(ald)
sseqa = seq(-5,5,0.5)
densa = dALD(y=sseqa,mu=0,sigma=1,p=0.5)
plot(sseqa,densa,type = "l",lwd=2,col="forestgreen",xlab="x",ylab="
    y", main="(a)")

densb = dALD(y=sseqa,mu=0,sigma=0.5,p=0.85)
plot(sseqa,densb,type = "l",lwd=2,col="forestgreen",xlab="x",ylab="
    y", main="(b)")

densc = dALD(y=sseqa,mu=0,sigma=0.5,p=0.15)
plot(sseqa,densc,type = "l",lwd=2,col="yellow3",xlab="x",ylab="y",
    main="(c)")
```

```
densd = dALD(y=sseqa,mu=2,sigma=0.5,p=0.5)
plot(sseqa,densd,type = "l",lwd=2,col="yellow3",xlab="x",ylab="y",
    main="(d)")

dense = dALD(y=sseqa,mu=0,sigma=2,p=0.5)
plot(sseqa,dense,type = "l",lwd=2,col="red",xlab="x",ylab="y", main
    ="(e)")

densf = dALD(y=sseqa,mu=0,sigma=0.2,p=0.5)
plot(sseqa,densf,type = "l",lwd=2,col="red",xlab="x",ylab="y", main
    ="(f)")

############## Quantile mixed-effects modelling ###########

library(car)
library(effects)
library(foreign)
library(lattice)
library(psych)
library(ggplot2)
library(papeR)

quant<- read.spss("Modified-Quantile.sav", to.data.frame = TRUE,
    use.missings = TRUE)

summary(quant)
hist(quant$sqrtCD4,ylab="Frequency",xlab="SQRT_CD4 cell count",main
    ="Histogram of CD4 count")
# descriptive statistics of CD4 by Baseline VL
tapply(quant$CD4_Count,quant$age_at_specimen_collection,mean)
tapply(quant$CD4_Count,quant$age_at_specimen_collection,sd)

# Or use describe() funtion from psych package
by(quant$CD4_Count,quant$age_at_baseline, describe)
by(quant$CD4_Count,quant$Baseline_VLoad, describe)

help(package = "qrLMM")
install.packages("qrLMM")
library(qrLMM)


quant$Educational_level<- as.factor(quant$Educational_level)
quant$Marital_status<- as.factor(quant$Marital_status)
```

```
quant$Residence_mod<- as.factor(quant$Residence_mod)
quant$ART<- as.factor(quant$ART)

attach(quant)
names(quant)
str(quant)
y=sqrtCD4 #response_SQRT_CD4_Count
x=cbind(1,TotalTimeMonth,SQRT_of_Time_month, Baseline_BMI1,
   Log_Baseline_VL,ART,Age) # design matrix for fixed effects
z=cbind(1,TotalTimeMonth,SQRT_of_Time_month)  #design matrix for
   random effects
fit.qmm<-QRLMM(y,x,z,group=PID,p=c(0.25,0.50,0.75),precision
   =0.0001,MaxIter=500,M=20,cp=0.25,
beta=NA,sigma=NA,Psi=NA,show.convergence=TRUE,CI=95)

### At single quantile ###
fit.qmm1<-QRLMM(y,x,z,group=PID,p= 0.95 ,precision=0.0001,MaxIter
   =500,M=20,cp=0.25,
beta=NA,sigma=NA,Psi=NA,show.convergence=TRUE,CI=95)
```

############################## **Chapter 5** ##############################

Code 7.4: R code for the data example in Section 5.6

```
GAMM<- read.spss("Modified-GAMM.sav", to.data.frame = TRUE, use.
   missings = TRUE)

str(GAMM)
summary(GAMM)
attach(GAMM)
names(GAMM)
head(GAMM)
tail(GAMM)

install.packages("mgcv")
install.packages("nlme")
library(nlme)
library(mgcv) #For GAM and GAMM

GAMM$Number_of_sextual_partner<- as.factor(
   GAMM$Number_of_sextual_partner)
GAMM$Highest_level_of_education<- as.factor(
   GAMM$Highest_level_of_education)
GAMM$Educational_level<- as.factor(GAMM$Educational_level)
```

```
GAMM$Marital_status<- as.factor(GAMM$Marital_status)
GAMM$Residence_mod<- as.factor(GAMM$Residence_mod)
GAMM$ART<- as.factor(GAMM$ART)
GAMM$PID<- as.factor(GAMM$PID)

### Modelling additive negative binomial mixed-effects model ###

gammNB<- gamm(CD4_Count~s(Age, k=20, bs="tp")+s(Time_in_Months, k
   =20, bs="tp")+s(Baseline_BMI1, k=20, bs="tp")+Educational_level+
   ART+Baseline_VL1+Residence_mod+Marital_status, family = nb ,
   method = "REML", random = list(PID=~1+Time_in_Months),
data = GAMM)  #Final model

summary(gammNB$gam)
summary(gammNB$lme)
gam.check(gammNB$gam)

plot(gammNB$lme, main="Standardized residuals")
plot(gamm$gam,shade = TRUE,shade.col = "palegreen",bty = "l")
anova(gammNB$gam)
intervals(gammNB$lme)
vis.gam(gammNB$gam)# 3-D plot
###To see plots in one group######
plot_numbers <- 1:4
layout(matrix(plot_numbers, ncol = 2, byrow = TRUE))
plot(gammNB$gam,shade = TRUE,shade.col = "palegreen",bty = "l",
   plot_numbers)
###plot of residuals versus fitted values###
diagnostics <- data.frame(
residuals = residuals(gammNB$gam),
fitted = fitted(gammNB$gam))
ggplot(diagnostics, aes(fitted, residuals)) +
geom_point() +
geom_smooth(method = "loess")
```

############################## **Chapter 6** ##############################

Code 7.5: R code for the data example in Section 6.5

```
To download and install Rtools-> https://cran.r-project.org/bin/
    windows/Rtools/
To install and compile Rtools4 -> writeLines('PATH="${RTOOLS40_HOME
    }\\usr\\bin;${PATH}"', con = "~/.Renviron")
Run this->writeLines('PATH="${RTOOLS40_HOME}\\usr\\bin;${PATH}"',
    con = "~/.Renviron")
Sys.which("make")
install.packages("devtools")
library(devtools)
devtools::install_github("marco-geraci/aqmm")
library(aqmm)
#install.packages("marco-geraci/aqmm")
install.packages("quantreg")

qamm<- read.spss("Modified- Quantile2- copy.sav", to.data.frame =
    TRUE, use.missings = TRUE)
summary(qamm)
attach(qamm)
# To get Geraci (2019) AQMM codes-> https://github.com/marco-geraci
    /aqmm/blob/master/man/aqmm.Rd
qamm$Educational_level<- as.factor(qamm$Educational_level)
qamm$Marital_status<- as.factor(qamm$Marital_status)
qamm$Residence_mod<- as.factor(qamm$Residence_mod)
qamm$Agegroup<- as.factor(qamm$Agegroup)
qamm$ART<- as.factor(qamm$ART)
qamm$PID<- as.factor(qamm$PID)

library(repmis)
library(quantreg)
library(dplyr)
library(lubridate)
library(stringr)
library(ggplot2)
library(usethis)
library(devtools)
library(splines)
fitB.05 <- rq(sqrtCD4 ~ bs(Baseline_BMI1, df=15), tau=0.05, data=
    qamm)
fitB.25 <- rq(sqrtCD4 ~ bs(Baseline_BMI1, df=15), tau=0.25, data=
    qamm)
fitB.50 <- rq(sqrtCD4 ~ bs(Baseline_BMI1, df=15), tau=0.50, data=
```

```
    qamm)
fitB.75 <- rq(sqrtCD4 ~ bs(Baseline_BMI1, df=15), tau=0.75, data=
    qamm)
fitB.85 <- rq(sqrtCD4 ~ bs(Baseline_BMI1, df=15), tau=0.85, data=
    qamm)
fitB.95 <- rq(sqrtCD4 ~ bs(Baseline_BMI1, df=15), tau=0.95, data=
    qamm)
fitB.99 <- rq(sqrtCD4 ~ bs(Baseline_BMI1, df=15), tau=0.99, data=
    qamm)

# Add quantiles to data frame

qamm<- qamm %>%
mutate(pc.99 = predict(fitB.99)) %>%
mutate(pc.95 = predict(fitB.95)) %>%
mutate(pc.85 = predict(fitB.85)) %>%
mutate(pc.75 = predict(fitB.75)) %>%
mutate(pc.50 = predict(fitB.50)) %>%
mutate(pc.25 = predict(fitB.25)) %>%
mutate(pc.05 = predict(fitB.05))

# plot
qamm %>%
ggplot(aes(x =Baseline_BMI1)) +
geom_point(aes(y = sqrtCD4)) +
geom_line(aes(y = pc.99, colour = '0.99 Quantile'))+
geom_line(aes(y = pc.95, colour = '0.95 Quantile'))+
geom_line(aes(y = pc.85, colour = '0.85 quantile'))+
geom_line(aes(y = pc.75, colour = '0.75 Quantile')) +
geom_line(aes(y = pc.50, colour = 'Median')) +
geom_line(aes(y = pc.25, colour = '0.25 Quantile')) +
geom_line(aes(y = pc.05, colour = '0.05 Quantile'))+
scale_color_manual('', values = c('0.99 Quantile' = 'blue', '0.95
    Quantile' = 'orange','0.85 Quantile' = 'red2','0.75 Quantile' =
    'red', 'Median' = 'yellow2', '0.25 Quantile' = 'green','0.05
    Quantile' = 'slategrey'),
breaks = c('0.99 Quantile','0.95 Quantile','0.85 Quantile','0.75
    Quantile', 'Median', '0.25 Quantile','0.05 Quantile')) +

xlab('Baseline BMI') +
ylab('SQRT CD4 Count')
########################################
str(qamm)
fitA.05 <- rq(sqrtCD4 ~ bs(Time_in_Months, df=15), tau=0.05, data=
```

```
    qamm)
fitA.25 <- rq(sqrtCD4 ~ bs(Time_in_Months, df=15), tau=0.25, data=
    qamm)
fitA.50 <- rq(sqrtCD4 ~ bs(Time_in_Months, df=15), tau=0.50, data=
    qamm)
fitA.75 <- rq(sqrtCD4 ~ bs(Time_in_Months, df=15), tau=0.75, data=
    qamm)
fitA.90 <- rq(sqrtCD4 ~ bs(Time_in_Months, df=15), tau=0.85, data=
    qamm)
fitA.95 <- rq(sqrtCD4 ~ bs(Time_in_Months, df=15), tau=0.95, data=
    qamm)
fitA.99 <- rq(sqrtCD4 ~ bs(Time_in_Months, df=15), tau=0.99, data=
    qamm)

# Add quantiles to data frame

qammOP<- qammOP %>%
mutate(pc.99 = predict(fitA.99)) %>%
mutate(pc.95 = predict(fitA.95)) %>%
mutate(pc.85 = predict(fitA.90)) %>%
mutate(pc.75 = predict(fitA.75)) %>%
mutate(pc.50 = predict(fitA.50)) %>%
mutate(pc.25 = predict(fitA.25)) %>%
mutate(pc.05 = predict(fitA.05))

# plot
qamm %>%
ggplot(aes(x =Time_in_Months)) +
geom_point(aes(y = sqrtCD4)) + ylim(5,47)+
geom_line(aes(y = pc.99, colour = '0.99 Quantile'))+
geom_line(aes(y = pc.95, colour = '0.95 Quantile'))+
geom_line(aes(y = pc.85, colour = '0.85 quantile'))+
geom_line(aes(y = pc.75, colour = '0.75 Quantile')) +
geom_line(aes(y = pc.50, colour = 'Median')) +
geom_line(aes(y = pc.25, colour = '0.25 Quantile')) +
geom_line(aes(y = pc.05, colour = '0.05 Quantile'))+
scale_color_manual('', values = c('0.99 Quantile' = 'blue','0.95
    Quantile' = 'orange','0.85 Quantile' = 'red2','0.75 Quantile' =
    'red', 'Median' = 'yellow2', '0.25 Quantile' = 'green','0.05
    Quantile' = 'slategrey'),
breaks = c('0.99 Quantile','0.95 Quantile','0.95 Quantile','0.75
    Quantile', 'Median', '0.25 Quantile','0.05 Quantile')) +

xlab('Treatment time (in months)') +
```

```
ylab('SQRT CD4 Count')

####The Model###
aqmm1<- aqmm(sqrtCD4~s(Time_in_Months, k=20, bs="tp")+s(
   Baseline_BMI1, k=20, bs="tp")+Age+Educational_level+ART+
   Baseline_VL1+Residence_mod+Marital_status, random=~1+
   Time_in_Months,
group = vistcode, knots = NULL, covariance = "pdDiag",data = qamm,
   tau = 0.05, gamm = TRUE, gradHess = FALSE, fit = TRUE) # tau =
   0.1, ..., 0.99
summary.aqmm(aqmm1)

#######For the predicted plot#####

aqmm1b<- aqmm(sqrtCD4~s(Time_in_Months, k=20, bs="tp")+s(
   Baseline_BMI1, k=20, bs="tp")+Age+Educational_level+ART+
   Baseline_VL1+Residence_mod+Marital_status, random=~1+
   Time_in_Months,
group = PID, knots = NULL, covariance = "pdDiag",data = qamm, tau =
    0.05, gamm = TRUE, gradHess = FALSE, fit = TRUE)

aqmm2b<- aqmm(sqrtCD4~s(Time_in_Months, k=20, bs="tp")+s(
   Baseline_BMI1, k=20, bs="tp")+Age+Educational_level+ART+
   Baseline_VL1+Residence_mod+Marital_status, random=~1+
   Time_in_Months,
group = PID, knots = NULL, covariance = "pdDiag",data = qamm, tau =
    0.25, gamm = TRUE, gradHess = FALSE, fit = TRUE)

aqmm3b<- aqmm(sqrtCD4~s(Time_in_Months, k=20, bs="tp")+s(
   Baseline_BMI1, k=20, bs="tp")+Age+Educational_level+ART+
   Baseline_VL1+Residence_mod+Marital_status, random=~1+
   Time_in_Months,
group = PID, knots = NULL, covariance = "pdDiag",data = qamm, tau =
    0.5, gamm = TRUE, gradHess = FALSE, fit = TRUE)

aqmm4b<- aqmm(sqrtCD4~s(Time_in_Months, k=20, bs="tp")+s(
   Baseline_BMI1, k=20, bs="tp")+Age+Educational_level+ART+
   Baseline_VL1+Residence_mod+Marital_status, random=~1+
   Time_in_Months,
group = PID, knots = NULL, covariance = "pdDiag",data = qamm, tau =
    0.75, gamm = TRUE, gradHess = FALSE, fit = TRUE)

aqmm5b<- aqmm(sqrtCD4~s(Time_in_Months, k=20, bs="tp")+s(
   Baseline_BMI1, k=20, bs="tp")+Age+Educational_level+ART+
```

```
    Baseline_VL1+Residence_mod+Marital_status, random=~1+
    Time_in_Months,
group = PID, knots = NULL, covariance = "pdDiag",data = qamm, tau =
     0.85, gamm = TRUE, gradHess = FALSE, fit = TRUE)

aqmm6b<- aqmm(sqrtCD4~s(Time_in_Months, k=20, bs="tp")+s(
    Baseline_BMI1, k=20, bs="tp")+Age+Educational_level+ART+
    Baseline_VL1+Residence_mod+Marital_status, random=~1+
    Time_in_Months,
group = PID, knots = NULL, covariance = "pdDiag",data = qamm, tau =
     0.95, gamm = TRUE, gradHess = FALSE, fit = TRUE)

aqmm7b<- aqmm(sqrtCD4~s(Time_in_Months, k=20, bs="tp")+s(
    Baseline_BMI1, k=20, bs="tp")+Age+Educational_level+ART+
    Baseline_VL1+Residence_mod+Marital_status, random=~1+
    Time_in_Months,
group = PID, knots = NULL, covariance = "pdDiag",data = qamm, tau =
     0.99, gamm = TRUE, gradHess = FALSE, fit = TRUE)

#######


plot(Time_in_Months, sqrtCD4, xlab="Treatment time (in months)",
    ylab="SQRT CD4 Count", ylim = range(15, 40), cex=0.0)

points(Time_in_Months, predict(aqmm1b), col = 'red', pch = 16, cex
    =0.0)

points(Time_in_Months, predict(aqmm2b), col = 'green', pch = 16,
    cex=0.0)

points(Time_in_Months, predict(aqmm3b), col = 'yellow2', pch = 16,
    cex=0.0)

points(Time_in_Months, predict(aqmm4b), col = 'black', pch = 16,
    cex=0.0)

points(Time_in_Months, predict(aqmm5b), col = 'red2', pch = 16, cex
    =0.0)

points(Time_in_Months, predict(aqmm6b), col = 'orange', pch = 16,
    cex=0.0)

points(Time_in_Months, predict(aqmm7b), col = 'blue', pch = 16, cex
```

```
    =0.0)


lines(loess.smooth(Time_in_Months, predict(aqmm1b), span =0.1),col
    ='slategrey', lwd=1, lty=1)
lines(loess.smooth(Time_in_Months, predict(aqmm2b), span =0.1),col
    ='green', lwd=1, lty=1)
lines(loess.smooth(Time_in_Months, predict(aqmm3b), span =0.1),col
    ='yellow2', lwd=1, lty=1)
lines(loess.smooth(Time_in_Months, predict(aqmm4b), span =0.1),col
    ='black', lwd=1, lty=1)
lines(loess.smooth(Time_in_Months, predict(aqmm5b), span =0.1),col
    ='red', lwd=1, lty=1)
lines(loess.smooth(Time_in_Months, predict(aqmm6b), span =0.1),col
    ='orange', lwd=1, lty=1)
lines(loess.smooth(Time_in_Months, predict(aqmm7b), span =0.1),col
    ='blue', lwd=1, lty=1)

legend("topleft",c("Quantiles", "0.05","0.25", "0.5","0.75",
    "0.85", "0.95", "0.99"),
col = c("gray90", "slategrey","green", "yellow2","black", "red","
    orange", "blue"),
cex = 0.4,text.col = "black",lty = c(1),lwd=c(1),pch = c(-1),
merge = TRUE, bg = 'gray90')
title(main= "Effect of treatment time")
####
plot(Baseline_BMI1, sqrtCD4, xlab="Baseline BMI", ylab="SQRT CD4
    Count", ylim = range(15, 42), cex=0.0)

points(Baseline_BMI1, predict(aqmm1b), col = 'red', pch = 16, cex
    =0.0)

points(Baseline_BMI1, predict(aqmm2b), col = 'green', pch = 16, cex
    =0.0)

points(Baseline_BMI1, predict(aqmm3b), col = 'yellow2', pch = 16,
    cex=0.0)

points(Baseline_BMI1, predict(aqmm4b), col = 'black', pch = 16, cex
    =0.0)

points(Baseline_BMI1, predict(aqmm5b), col = 'red2', pch = 16, cex
    =0.0)
```

```
points(Baseline_BMI1, predict(aqmm6b), col = 'orange', pch = 16,
    cex=0.0)

points(Baseline_BMI1, predict(aqmm7b), col = 'blue', pch = 16, cex
    =0.0)


lines(loess.smooth(Baseline_BMI1, predict(aqmm1b), span =0.1),col='
    slategrey', lwd=1, lty=1)
lines(loess.smooth(Baseline_BMI1, predict(aqmm2b), span =0.1),col='
    green', lwd=1, lty=1)
lines(loess.smooth(Baseline_BMI1, predict(aqmm3b), span =0.1),col='
    yellow2', lwd=1, lty=1)
lines(loess.smooth(Baseline_BMI1, predict(aqmm4b), span =0.1),col='
    black', lwd=1, lty=1)
lines(loess.smooth(Baseline_BMI1, predict(aqmm5b), span =0.1),col='
    red', lwd=1, lty=1)
lines(loess.smooth(Baseline_BMI1, predict(aqmm6b), span =0.1),col='
    orange', lwd=1, lty=1)
lines(loess.smooth(Baseline_BMI1, predict(aqmm7b), span =0.1),col='
    blue', lwd=1, lty=1)

legend("topleft",c("Quantiles", "0.05","0.25", "0.5","0.75",
    "0.85", "0.95", "0.99"),
col = c("gray90", "slategrey","green", "yellow2","black", "red","
    orange", "blue"),
cex = 0.4,text.col = "black",lty = c(1),lwd=c(1),pch = c(-1),
merge = TRUE, bg = 'gray90')
title(main = "Effect of baseline BMI")
```

Appendix B: Additional Results

Chapter 3: Additional outputs

Table 7.1: Comparison of covariance structure using the fitted model (Model 1)

		Information Criteria				
Covariance Structure	$-2\log\ell$	**AIC**	**AICC**	**BIC**	**CAIC**	**HQIC**
AR(1)	89116.94	89116.94	89161.09	89237.05	89259.05	89191.63
CS	89135.76	89179.76	89179.91	89255.87	89277.87	89210.45
Toep	89113.46	89159.46	89159.62	89239.03	89262.03	89191.54
UN	**87781.28**	**87833.28**	**87833.48**	**87923.23**	**87949.23**	**87869.54**
VC	88069.85	88115.85	88116.00	88195.42	88218.42	88147.93
ARH(1)	87968.69	88016.69	88016.86	88099.72	88123.72	88050.17
CSH	87893.60	87941.60	87941.78	88024.63	88048.63	87975.08
ToepH	87888.63	87938.63	87938.81	88025.12	88050.12	87973.50

Table 7.2: Unstructured covariance Parameter Estimates

Cov Parm	**Subject**	**Estimate**
UN(1,1)	PID	0.1131
UN(2,1)	PID	0.000739
UN(2,2)	PID	0.000155
UN(3,1)	PID	-0.01754
UN(3,2)	PID	-0.00137
UN(3,3)	PID	0.01556
Scale		0.04205

Table 7.3: Comparison of fixed effects results across different covariance structure using Model 1

	UN		AR(1)		CS		Toep	
Covariates	**Estimate**	**Std Err**	**Estimate**	**Std Err**	**Estimate**	**Std Err**	**Estimate**	**Std Err**
Intercept	6.4697	0.04982	6.4724	0.03423	6.4861	0.03410	6.4799	0.03439
Time in month	0.007824	0.000989	0.008516	0.01060	0.01439	0.01051	0.008272	0.01082
Sqrt_Time	-0.08649	0.009307	-0.08950	0.01180	-0.08434	0.01170	-0.08886	0.01201
ART Initiation (Post)	0.2301	0.01238	0.2284	0.01263	0.2363	0.01265	0.2277	0.01264
Baseline BMI category								
Obese	0.4815	0.1113	0.6076	0.07836	0.5097	0.07765	0.6350	0.07813
Overweight	0.02561	0.04975	0.02687	0.03466	0.02072	0.03441	0.02970	0.03448
Underweight	0.005901	0.07927	0.09673	0.05503	0.03837	0.05470	0.09359	0.05481
Baseline HIV VL category								
High VL	-0.2393	0.05157	-0.3307	0.03345	-0.3234	0.03321	-0.3377	0.03330
Medium VL	-0.1258	0.04587	-0.1527	0.03130	-0.1254	0.03112	-0.1567	0.03116
Undetectable	0.1377	0.2901	-0.04788	0.2242	0.1338	0.2256	-0.01985	0.2218
Number of sex partner								
Many partners	-0.1560	0.09394	-0.05213	0.06388	-0.1506	0.06352	-0.04274	0.06393
No partner	-0.04821	0.04993	-0.03423	0.03459	-0.05490	0.03434	-0.03322	0.03438
Age group in years								
20-29	0.01166	0.03104	0.02553	0.02516	0.006652	0.02519	0.02065	0.02543
30-39	0.02852	0.03432	0.04911	0.02849	0.03351	0.02850	0.04303	0.02871
40-49	-0.00719	0.04545	0.007849	0.04070	0.01926	0.04068	-0.00114	0.04084
50-59	-0.05694	0.06662	-0.06551	0.06134	-0.03957	0.06135	-0.06503	0.06143
≥ 60	0.2082	0.1532	-0.2185	0.1606	0.2020	0.1601	-0.1844	0.1612
Education level								
Primary school	-0.04509	0.09084	0.1126	0.06341	-0.00666	0.06299	0.09430	0.06306
Residence of participant								
Rural	-0.00373	0.03947	0.003881	0.02707	0.01729	0.02689	0.003076	0.02694

- The reference categories are the same as in Table 3.7.

Chapter 4: Additional outputs

R package *qrLMM()* sample output using CAPRISA 002 AI Study dataset

```
--------------------------------------------------
Quantile Regression for Linear Mixed Model
--------------------------------------------------
Quantile = 0.5
Subjects = 235; Observations = 7019
-----------
Estimates
-----------
- Fixed effects

       Estimate Std. Error Inf CI95% Sup CI95%  z value Pr(>|z|)
beta 1 24.62849    1.46389  21.75927  27.49770 16.82404  0.00000
beta 2  0.05678    0.01285   0.03159   0.08197  4.41807  0.00001
beta 3 -0.69589    0.11792  -0.92702  -0.46477 -5.90142  0.00000
beta 4  0.08170    0.02699   0.02879   0.13461  3.02638  0.00248
beta 5 -0.64095    0.09631  -0.82972  -0.45219 -6.65513  0.00000
beta 6  2.56004    0.08808   2.38740   2.73268 29.06490  0.00000
beta 7  0.02935    0.03125  -0.03191   0.09060  0.93898  0.34774

sigma = 0.87713

Random effects Variance-Covariance Matrix
         z1       z2       z3
z1 17.69164  0.16071 -2.93488
z2  0.16071  0.02010 -0.18231
z3 -2.93488 -0.18231  2.03916

------------------------
Model selection criteria
------------------------
         Loglik      AIC      BIC       HQ
Value -16828.96 33685.92 33781.91 33718.99
-------
Details
-------
Convergence reached? = TRUE
Iterations = 433 / 500
Criteria = 0.00036
MC sample = 20
Cut point = 0.25
Processing time = 2.591965 hours
```

Table 7.4: Parameter estimates at 0.05^{th} quantile

Parameter	Estimate	St.error	95% C.I	P-value
Intercept	19.99675	1.16072	(17.72173 22.27177)	0.00000
Time	0.06294	0.01473	(0.03407 0.09180)	0.00002
SQRT of Time	-0.86567	0.14159	(-1.14319 -0.58815)	0.00000
Baseline BMI	0.05642	0.02067	(0.01590 0.09694)	0.00635
Log of baseline VL	-0.56383	0.07799	(-0.71670 -0.41096)	0.00000
Post HAART initiation	1.68287	0.05379	(1.57744 1.78830)	0.00000
Age	0.02073	0.02500	(-0.02826 0.06972)	0.40688

Table 7.5: Parameter estimates at 0.25^{th} quantile

Parameter	Estimate	St.error	95% C.I	P-value
Intercept	22.17136	1.40335	(19.42079 24.92193)	0.00000
Time	0.06979	0.01369	(0.04296 0.09661)	0.00000
SQRT of Time	-0.87112	0.12946	(-1.12486 -0.61738)	0.00000
Baseline BMI	0.07836	0.02432	(0.03070 0.12602)	0.00127
Log of baseline VL	-0.56874	0.10345	(-0.77149 -0.36598)	0.00000
Post HAART initiation	2.12541	0.07329	(1.98176 2.26907)	0.00000
Age	0.02957	0.02972	(-0.02868 0.08781)	0.31975

Table 7.6: Parameter estimates at 0.85^{th} quantile

Parameter	Estimate	St.error	95% C.I	P-value
Intercept	27.97228	1.42043	(25.18824 30.75631)	0.00000
Time	0.04132	0.01313	(0.01559 0.06705)	0.00165
SQRT of Time	-0.58146	0.12470	(-0.82587 -0.33705)	0.00000
Baseline BMI	0.13131	0.03383	(0.06500 0.19762)	0.00010
Log of baseline VL	-0.71471	0.08997	(-0.89105 -0.53837)	0.00000
Post HAART initiation	3.11409	0.09728	(2.92342 3.30476)	0.00000
Age	0.02576	0.03215	(-0.03725 0.08878)	0.42294

Table 7.7: Parameter estimates at 0.95^{th} quantile

Parameter	Estimate	St.error	95% C.I	P-value
Intercept	31.38118	1.39665	(28.64373 34.11862)	0.00000
Time	0.03366	0.01578	(0.00273 0.06459)	0.03293
SQRT of Time	-0.38521	0.16218	(-0.70309 -0.06734)	0.01754
Baseline BMI	0.14515	0.03043	(0.08551 0.20480)	0.00000
Log of baseline VL	-0.73982	0.08494	(-0.90631 -0.57333)	0.00000
Post HAART initiation	2.28722	0.08890	(2.11298 2.46146)	0.00000
Age	0.01328	0.03032	(-0.04615 0.07271)	0.66130

Chapter 6: Additional outputs

Table 7.8: R package additive quantile mixed model, aqmm(), sample outputs using CAPRISA 002 Acute Infection Study data across various quantile levels

```
Quantile 0.25

Fixed effects:
                                  Estimate  Std. Error t value  Pr(>|t|)
(Intercept)                     1.9647e+01  4.7491e-01 41.3708 < 2.2e-16 ***
Age                             2.0960e-02  1.1454e-02  1.8299  0.068760 .
Educational_levelSecondary school -4.7338e-01  4.1009e-01 -1.1543  0.249745
ARTpost ART initiation           1.5296e+00  5.9779e-02 25.5868 < 2.2e-16 ***
Baseline_VL1                    -2.0864e-06  2.6911e-07 -7.7530 4.538e-13 ***
Residence_mod2                   2.4989e-01  5.4473e-02  4.5875 7.927e-06 ***
Marital_statusStable             3.0457e-01  1.5489e-01  1.9664  0.050647 .
Marital_statusMany              -7.8584e-01  2.5890e-01 -3.0353  0.002724 **
s(Time_in_Months)Fx1            -2.3766e+00  5.5497e-01 -4.2824 2.875e-05 ***
s(Baseline_BMI1)Fx1              5.6868e+00  1.1094e+00  5.1261 6.982e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Covariance matrix of the random effects:
              (Intercept) Time_in_Months
(Intercept)    8.687e-01   0.000e+00
Time_in_Months 0.000e+00   1.929e-16

Variances of the (random) smooth terms:
s(Time_in_Months)  s(Baseline_BMI1)
            28.94           6463.83

Residual scale parameter:  1.335
Log-likelihood: -21031
Tuning parameter: 0.0003077

Number of observations: 7019
Number of groups: 235
```

```
Quantile 0.75

Fixed effects:
                                    Estimate  Std. Error t value  Pr(>|t|)
(Intercept)                       2.4167e+01  1.0536e+00 22.9374 < 2.2e-16 ***
Age                               3.3143e-02  7.7590e-03  4.2716 3.006e-05 ***
Educational_levelSecondary school 3.8519e-01  1.0677e+00  0.3608  0.718663
ARTpost ART initiation            1.5292e+00  5.4576e-02 28.0196 < 2.2e-16 ***
Baseline_VL1                     -1.5700e-06  1.6001e-07 -9.8123 < 2.2e-16 ***
Residence_mod2                    1.2747e-01  1.4362e-01  0.8876  0.375851
Marital_statusStable              4.9069e-01  1.5944e-01  3.0776  0.002381 **
Marital_statusMany               -1.1719e+00  2.5698e-01 -4.5603 8.916e-06 ***
s(Time_in_Months)Fx1             -2.2829e+00  4.9993e-01 -4.5665 8.680e-06 ***
s(Baseline_BMI1)Fx1               5.7904e+00  1.2077e+00  4.7945 3.188e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Covariance matrix of the random effects:
              (Intercept) Time_in_Months
(Intercept)    2.453e-01   0.000e+00
Time_in_Months 0.000e+00   5.451e-17

Variances of the (random) smooth terms:
s(Time_in_Months)  s(Baseline_BMI1)
            30.28           6290.32

Residual scale parameter:  1.43
Log-likelihood: -21680
Tuning parameter: 0.0003077

Number of observations: 7019
Number of groups: 235
```

```
Quantile 0.85

Fixed effects:
                                    Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)                       2.5845e+01  8.8147e-01  29.3207 < 2.2e-16 ***
Age                               3.1118e-02  1.2061e-02   2.5802 0.0105958 *
Educational_levelSecondary school 5.0743e-01  9.6045e-01   0.5283 0.5978665
ARTpost ART initiation            1.4546e+00  1.2778e-01  11.3838 < 2.2e-16 ***
Baseline_VL1                     -1.7132e-06  1.5806e-07 -10.8391 < 2.2e-16 ***
Residence_mod2                   -6.0903e-02  1.9103e-01  -0.3188 0.7501999
Marital_statusStable              6.2981e-01  1.3699e-01   4.5974 7.594e-06 ***
Marital_statusMany               -1.5711e+00  4.2531e-01  -3.6940 0.0002852 ***
s(Time_in_Months)Fx1             -2.2924e+00  3.8936e-01  -5.8877 1.642e-08 ***
s(Baseline_BMI1)Fx1               5.6964e+00  1.1277e+00   5.0513 9.898e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Covariance matrix of the random effects:
              (Intercept) Time_in_Months
(Intercept)    3.454e-01   0.000e+00
Time_in_Months 0.000e+00   7.671e-17

Variances of the (random) smooth terms:
s(Time_in_Months)  s(Baseline_BMI1)
            21.92           4979.39

Residual scale parameter:  1.068
Log-likelihood: -22972
Tuning parameter: 0.0002564

Number of observations: 7019
Number of groups: 235
```

```
Quantile 0.99

Fixed effects:
                                    Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)                       2.9479e+01  7.2187e-01  40.8373 < 2.2e-16 ***
Age                               3.5241e-02  3.2735e-02   1.0765 0.2829879
Educational_levelSecondary school 2.0311e+00  9.3341e-01   2.1760 0.0307302 *
ARTpost ART initiation            2.8819e+00  2.1028e-01  13.7050 < 2.2e-16 ***
Baseline_VL1                     -1.6168e-06  2.1655e-07  -7.4661 2.531e-12 ***
Residence_mod2                   -1.4498e+00  3.0687e-01  -4.7243 4.356e-06 ***
Marital_statusStable              2.8311e+00  1.8621e-01  15.2040 < 2.2e-16 ***
Marital_statusMany               -4.2155e+00  1.6546e-01 -25.4775 < 2.2e-16 ***
s(Time_in_Months)Fx1             -1.6920e+00  4.6815e-01  -3.6143 0.0003815 ***
s(Baseline_BMI1)Fx1               5.2328e+00  1.1256e+00   4.6487 6.075e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Covariance matrix of the random effects:
               (Intercept) Time_in_Months
(Intercept)    3.326e-03   0.000e+00
Time_in_Months 0.000e+00   2.963e-18

Variances of the (random) smooth terms:
s(Time_in_Months)  s(Baseline_BMI1)
            2.669           576.902

Residual scale parameter:  0.1276
Log-likelihood: -30865
Tuning parameter: 0.0002404

Number of observations: 7019
Number of groups: 235
```

Appendix C: Supplementary Materials

C.1 Important Determinants and Inverse properties for differentiation of matrix expression

Determinants

Suppose $\boldsymbol{A}$ is a square matrix having elements that are not functionally related. Then denoting the cofactor of a_{ij} in $|\boldsymbol{A}|$ by $|\boldsymbol{A}_{ij}|$, we have

$$\frac{\partial|\boldsymbol{A}|}{\partial a_{ij}} = |\boldsymbol{A}_{ij}|$$

One particular case of which is

$$\frac{\partial|\boldsymbol{A}|}{\partial a_{ij}} = |\boldsymbol{A}_{ji}|, \text{ iff } |\boldsymbol{A}| \text{ is symmetric, not for } i \neq j.$$

Then in place of the above expression, we have

$$\begin{aligned}\frac{\partial|\boldsymbol{A}|}{\partial \boldsymbol{\theta}} &= \frac{\partial|\boldsymbol{A}|}{\partial a_{ij}}\frac{\partial a_{ij}}{\partial \theta} + \frac{\partial|\boldsymbol{A}|}{\partial a_{ji}}\frac{\partial a_{ji}}{\partial \theta} \\ &= |\boldsymbol{A}_{ij}| + |\boldsymbol{A}_{ji}| \\ &= 2|\boldsymbol{A}_{ij}|,\end{aligned}$$

because $\boldsymbol{A}$ is symmetric. Hence, in general

$$\frac{\partial|\boldsymbol{A}|}{\partial a_{ij}} = (2 - \delta_{ij}|\boldsymbol{A}_{ij}|) \text{ for symmetric } \boldsymbol{A},$$

where δ_{ij} is the Kronecker delta, $\delta_{ij} = 0$ for $i \neq j$ and $\delta_{ij} = 1$ for $i = j$ (Searle, 1982).

Suppose that elements of $\boldsymbol{A}$ are functions of scalar t. Then,

$$\begin{aligned}\frac{\partial \log|\boldsymbol{A}|}{\partial t} &= \frac{1}{|\boldsymbol{A}|}\frac{\partial|\boldsymbol{A}|}{\partial t} = \frac{1}{|\boldsymbol{A}|}\sum\sum_{i\le j}\frac{\partial|\boldsymbol{A}|}{\partial a_{ij}}\frac{\partial a_{ij}}{\partial t} \\ &= \frac{1}{|\boldsymbol{A}|}\sum\sum_{i\le j}(2-\delta_{ij})|\boldsymbol{A}_{ij}|\frac{\partial a_{ij}}{\partial t} \\ &= \frac{1}{|\boldsymbol{A}|}\sum_i\sum_j|\boldsymbol{A}_{ij}|\frac{\partial a_{ij}}{\partial t} = \sum_i\sum_j\frac{|\boldsymbol{A}_{ij}|}{|\boldsymbol{A}|}\frac{\partial a_{ij}}{\partial t} \\ &= \sum_i\sum_j a^{ij}\frac{\partial a_{ij}}{\partial t} = tr\left[(\boldsymbol{A}^{-1})'\frac{\partial \boldsymbol{A}}{\partial t}\right] \\ &= tr\left(\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial t}\right), \qquad for \quad \boldsymbol{A}^{-1}=\{a^{ij}\}\end{aligned}$$

This result is used in deriving maximum likelihood equations for estimating variance components in Section 2.4.

Inverse

The *Inverse* of a square matrix $\boldsymbol{A}$, denoted $\boldsymbol{A}^{-1}$, is defined as a square matrix whose elements follow the following property:

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I},$$

where $\boldsymbol{I}$ is the identity matrix. With scalar t, we define
$\frac{\partial \boldsymbol{A}}{\partial t} = \left\{\frac{\partial a_{ij}}{\partial t}\right\}$
with $\boldsymbol{A}$ non-singular, $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}$ gives
$\frac{\partial \boldsymbol{A}}{\partial t}\boldsymbol{A}^{-1} + \boldsymbol{A}\frac{\partial \boldsymbol{A}^{-1}}{\partial t} = 0$. Therefore,
$\frac{\partial \boldsymbol{A}^{-1}}{\partial t} = -\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial t}\boldsymbol{A}^{-1}$

C.2 Vectors orthogonal to columns of X (for REML equation)

Suppose the set of values $\boldsymbol{k'Y}$ is chosen such that no term in the fixed effects are contained $\boldsymbol{k'x} = 0$. Then $\boldsymbol{x'k} = 0$ and, from the theory of solving linear equations (Searle, 1982), $\boldsymbol{k} = [\boldsymbol{I} - (\boldsymbol{x}')^{-}\boldsymbol{x}']\boldsymbol{c}$ for any vector $\boldsymbol{c}$, of appropriate order. Therefore, since $(\boldsymbol{x}^{-})'$ is a generalized inverse of $\boldsymbol{x}'$ we can write $\boldsymbol{k}' = \boldsymbol{c}'(\boldsymbol{I} - \boldsymbol{x}\boldsymbol{x}^{-})$. Moreover, because $(\boldsymbol{x'x})^{-}\boldsymbol{x}'$ is a generalized inverse of $\boldsymbol{x}$ another form for $\boldsymbol{k}'$ is $\boldsymbol{k}' = \boldsymbol{c}'[\boldsymbol{I} - \boldsymbol{x}(\boldsymbol{x'x})^{-}\boldsymbol{x}']$; as is $\boldsymbol{c}'(\boldsymbol{I} - \boldsymbol{x}\boldsymbol{x}^{+})$ since $\boldsymbol{x}(\boldsymbol{x'x})^{-}\boldsymbol{x}' = \boldsymbol{x}\boldsymbol{x}^{+}$. Thus, the two forms of $\boldsymbol{k}'$ are

$$\boldsymbol{x}' = \boldsymbol{c}'(\boldsymbol{I} - \boldsymbol{x}\boldsymbol{x}^{-}) \quad \text{or} \quad \boldsymbol{x}' = \boldsymbol{c}'[\boldsymbol{I} - \boldsymbol{x}(\boldsymbol{x'x})^{-}\boldsymbol{x}'] = \boldsymbol{c}'(\boldsymbol{I} - \boldsymbol{x}\boldsymbol{x}^{+}).$$

With $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{x}(\boldsymbol{x'x})^{-}\boldsymbol{x}' = \boldsymbol{I} - \boldsymbol{x}\boldsymbol{x}^{+}$, we therefore have $\boldsymbol{k}' = \boldsymbol{c}'\boldsymbol{M}$. With $\boldsymbol{x}$ of order $N \times p$ of rank r, there are only $N - r$ linearly independent vectors $\boldsymbol{k}'$ satisfying $\boldsymbol{k'x} = 0$ (Searle, 1982). Using a set of such $N - r$ linearly independent vectors $\boldsymbol{k}'$

as rows of $\boldsymbol{K}'$, we then have the following theorem, for $\boldsymbol{k'x} = 0$ with $\boldsymbol{K'}$ having maximum row rank $N - r$ and $\boldsymbol{K}' = \boldsymbol{C'M}$ for some $\boldsymbol{C}$.

C.3 Cholesky decomposition

The classical approach to eliminate all sources of correlation by appropriate scaling to check residuals in linear mixed models is applying the Cholesky decomposition for the generation of transformed residuals that have constant variance and zero correlation. The application of the Cholesky decomposition on the variance-covariance matrix starts with construction of a lower triangular matrix for each subject, denoted by $\hat{\boldsymbol{C}}_i$, which satisfies the condition:

$$\hat{\boldsymbol{V}}_i = \hat{\boldsymbol{C}}_i \hat{\boldsymbol{C}}_i',$$

where $\hat{\boldsymbol{C}}_i$ represents the Cholesky root of $\hat{\boldsymbol{V}}_i$, with $\hat{\boldsymbol{C}}_i^{-1}\boldsymbol{Y}_i$ having constant variance and zero correlation Liu (2015).

Given the attached properties, the $\hat{\boldsymbol{C}}_i$ matrix can be used to transform the correlated residuals to correlation free transformed residuals. For the marginal distribution of longitudinal data, the scaled residuals, denoted by $\boldsymbol{R}_{m_i}$, are defined as

$$\boldsymbol{R}_{m_i} = \hat{\boldsymbol{C}}_i^{-1}\boldsymbol{r}_{m_i} = \hat{\boldsymbol{C}}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i'\hat{\boldsymbol{\beta}}),$$

which have unit variance and zero correlation.

In mathematical form, if $\boldsymbol{A}$ is a symmetric positive definite matrix, then there exists an upper triangular matrix $\boldsymbol{C}$ such that

$$\boldsymbol{A} = \boldsymbol{CC}'.$$

The right-hand side of the above equation is called the *Cholesky decomposition* of the matrix $\boldsymbol{A}$.

An example of a *Cholesky decomposition* is

$$\begin{bmatrix} 4 & 10 \\ 10 & 169 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 5 & 12 \end{bmatrix} \begin{bmatrix} 2 & 5 \\ 0 & 12 \end{bmatrix}$$

- There exist fast and numerically stable algorithms for computing the *Cholesky decomposition*, which is an important tool for matrix computations (Ruppert et al., 2003; Gentle, 2009)

C.4 Quantiles as solutions of a minimization problem

Let Y be a continuous random variable, the expected value of the absolute sum of deviations from a given center c can be split into the following two terms:

$$\begin{aligned} E|Y-c| &= \int_{y\in R} |Y-c| f(y)dy \\ &= \int_{y<c} |Y-c| f(y)dy + \int_{y>c} f(y)dy \\ &= \int_{y<c} (c-Y) f(y)dy + \int_{y>c} (Y-c) f(y)dy \end{aligned}$$

Since the absolute value is a convex function, differentiating $E|Y-c|$ with respect to c and setting the partial derivatives to zero will lead to the solution for the minimum:

$$\frac{\partial}{\partial c} E|Y-c|.$$

The solution to the above equation can be obtained applying the derivative and integrating per part as presented in Davino et al. (2013).

C.5 Gibbs Sampling

Gibbs sampling is a Markovian (Markov Chain) algorithm introduced by Geman & Geman (1984) and has been mainly applied in the context of complex stochastic models involving very large numbers of variables. Given an arbitrary starting set of values $U_1^{(0)}, U_2^{(0)}, ..., U_k^{(0)}$, we can draw $U_1^{(1)} \sim \left[U_1 | U_2^{(0)}, U_3^{(0)}, ..., U_k^{(0)}\right]$ then $U_2^{(1)} \sim \left[U_2 | U_1^{(1)}, U_3^{(0)}, U_4^{(0)}, ..., U_k^{(0)}\right]$, $U_3^{(1)} \sim \left[U_3 | U_1^{(1)}, U_2^{(1)}, U_4^{(0)}, U_5^{(0)}, ..., U_k^{(0)}\right]$, and so on, up to $U_k^{(1)} \sim \left[U_k | U_1^{(1)}, U_2^{(1)}, ..., U_{k-1}^{(1)}\right]$, which is known as Gibbs sampling. Thus, each variable is *visited* in the *natural* order and a cycle in this scheme requires k random variate generations. After i such iterations we would arrive at $(U_1^{(i)}, ..., U_k^{(i)})$. See Gelfand & Smith (1990), Gelfand et al. (1990), Geman & Geman (1984), Casella & George (1992), Gelfand (2000), and the references therein, for further discussion.

Recently, Galarza et al. (2015) described the procedure of Gibbs sampler for obtaining a sequence of observations which are approximated from the joint probability distribution of two or more random variables using their full conditional distribution. One can refer to that reference for more details.

Appendix D: Published Papers

Modelling CD4 counts before and after HAART for HIV infected patients in KwaZulu-Natal South Africa

Ashenfai A Yirga[1], Sileshi F Melesse[1], Henry G Mwambi[1], Dawit G Ayele[2]

1. School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, Private Bag X01, Scottsville, 3209, South Africa.
2. Institute of Human Virology, University of Maryland, School of Medicine, USA.

Abstract

Background: This study aims to make use of a longitudinal data modelling approach to analyze data on the number of CD4+cell counts measured repeatedly in HIV-1 Subtype C infected women enrolled in the Acute Infection Study of the Centre for the AIDS Programme of Research in South Africa.

Methodology: This study uses data from the CAPRISA 002 Acute Infection Study, which was conducted in South Africa. This cohort study observed N=235 incident HIV-1 positive women whose disease biomarkers were measured repeatedly at least four times on each participant.

Results: From the findings of this study, post-HAART initiation, baseline viral load, and the prevalence of obese nutrition status were found to be major significant factors on the prognosis CD4+ count of HIV-infected patients.

Conclusion: Effective HAART initiation immediately after HIV exposure is necessary to suppress the increase of viral loads to induce potential ART benefits that accrue over time. The data showed evidence of strong individual-specific effects on the evolution of CD4+ counts. Effective monitoring and modelling of disease biomarkers are essential to help inform methods that can be put in place to suppress viral loads for maximum ART benefits that can be accrued over time at an individual level.

Keywords: Random-effects model; spatial covariance structure; CD4+ count; HAART; CAPRISA.

DOI: https://dx.doi.org/10.4314/ahs.v20i4.7

Cite as: *Yirga AA, Melesse SF, Mwambi HG, Ayele DG. Modelling CD4 counts before and after HAART for HIV infected patients in KwaZulu-Natal South Africa. Afri Health Sci. 2020;20(4):1546-61. https://dx.doi.org/10.4314/ahs.v20i4.7*

Background

Multilevel data modelling allows to account for the correlation of measurements, and include variables measured at different levels as well as model the variation at different levels. Longitudinal data, or repeated measurements data is a specific form of multilevel data. In longitudinal studies, repeated observations are made on an individual on one or more outcomes, including covariate information at a baseline and over time. Measurements made on the same individual are likely to be more similar than measurements made on different individuals. Thus, observations on the same individual will not be independent. That is, repeated measurements on the same subjects are bound to be correlated [1-3].

Longitudinal data analysis is widely used for at least three reasons: to increase the sensitivity by making within-subject comparisons, to study changes over time, and to use subjects efficiently once they are enrolled in a study[4-6]. Repeated measurements can compensate for small sample sizes because an individual is observed more than once compared to a cross-sectional study. The need for the covariance structure of the observed data makes longitudinal data analysis more complex than standard linear regression. For the inference to be substantial, the covariance among repeated measures must be appropriately modeled. Although the covariance structure is not the prime interest of the study, it is essential for valid inference [7,8]. Therefore, a lot of efforts are needed at the beginning of the statistical analysis to assess the covariance structure of the data. Traditional methods for longitudinal data such as Analysis of Variance (ANOVA) and Multivariate Analysis

Corresponding author:
Dawit G Ayele,
Institute of Human Virology,
University of Maryland,
School of Medicine, USA.
Email: ejigmul@yahoo.com, ejigmul@gmail.com

of Variance (MANOVA) are of limited use because of the restrictive assumptions concerning the variance-covariance structure of the repeated measures [9]. For this reason, mixed-effects models have become popular for modelling longitudinal data. This statistical procedure also permits the estimation of variability in hierarchically structured data and examines the impacts of factors at distinctive levels [10,11]. Since longitudinal studies are often faced with the incompleteness of the data due to partially observed subjects, the mixed-effects model is by its very nature able to deal with unbalanced data of this nature.

Thus, this study was conducted to review the general Linear Mixed Model approach that can be extended for multivariate longitudinal data by assuming appropriate random effects. This method has the benefit of having extra correlation evolving from the longitudinal data structure that can be modeled within the same framework. Therefore, the focus of this study is to adopt the mixed-effects model with appropriate random effects incorporated, including a flexible variance-covariance structure that gives the best fit as well as identifying whether specific clinical and sociodemographic factors present in the data (and their respective possible interactions) influenced CD4 count in a cohort of HIV-Infected Patients. The information and understanding of such factors are of epidemiological importance. The results will be beneficial in developing tools to support clinicians in the identification of factors related to HIV-Infected Patients. The results can be further used to shape communication and counseling strategies at the individual level before treatment initiation.

Materials and methods

Data source: This study uses data from the Centre for the AIDS Programme of Research in South Africa (CAPRISA) 002 Acute Infection Study. The study was conducted on HIV-infected women at the Doris Duke Medical Research Institute (DDMRI) at the Nelson R Mandela School of Medicine of the University of KwaZulu-Natal in Durban, South Africa. Between August 2004 and May 2005, CAPRISA initiated a cohort study enrolling high-risk HIV negative women to follow up. Women infected with HIV were recruited into the Acute Infection Study and then followed up closely to study disease progression and CD4/viral load evolution [12-14]. Once HIV-Infected women enrolled in the AI study, their CD4 cell count and viral load were measured and assessed regularly. When their CD4 cell count is less than or equal to 350 cells/mm^3 for more than two consecutive visits between 6 months or if they were with AIDS-defining illness (WHO clinical stage 3-5), they would be referred to a public government clinic for ARV treatment. However, these patients would only start HAART once their CD4 cell count was less than 200 cells/mm^3, according to the National Department of Health South Africa until 2015. With effect from the 1st January 2015, according to the National Department of Health, the criterion to start HIV patients on early initiation of ART was a CD4 cell count less than or equal to 500 cells/mm^3 [32,33].

Method

Mixed-effects modelling is an advanced and vital method in statistics. It is a well-known method; therefore, we summarize the key aspects of the model relevant to the current study. The literature on mixed models is ubiquitous, and some of it can be found in [2,3,5,6,9,11,15-18].

The use of the mixed-effects model for longitudinal data helps to correctly account for the correlation of observations within a subject and also to quantify the heterogeneity between subjects due to unobserved factors. It is important that before its implementation, adequate sample size is determined based on prior information on the magnitude of the correlation and the planned number of observations per individual. By correctly estimating the sample size, we end up with correctly estimated standard errors (SEs), which will give reliable confidence intervals (CI) and p-values. We can use the mixed-effects model to account for variation at lower and higher levels of the design structure. Accounting for variation at a lower level gives us more power for estimation at a higher level [3]. A mixed model is made up of fixed and random effects where the latter helps in accounting for correlation at a lower level within higher-level units. That is why mixed models are called "mixed" because the coefficients are a mix of fixed and random effects.

In more general terms, fixed effects or fixed factors are covariates that we anticipate will influence the outcome variable. They are what we call explanatory variables in a standard linear regression. For instance, in our case, we are interested in making conclusions about how the socio-economic, demographic, and treatment type (place of residence, baseline BMI, baseline viral load, age, education level, marital status, HAART initiation, etc.) impacts the CD4+ count of a patient. Therefore, these socio-economic, demographic, and treatment types are fixed effects, and CD4+ count of a patient is the response variable. Thus, a fixed-effect is the param-

eter of interest. The overall intercept is not the variable of interest, but of course, it is a fixed effect. In addition to the fixed effects, we also incorporate random effects in the mixed-effect model. Random effects are grouping factors for which we are attempting to control. A random intercept allows a different intercept for each subject. A random effect for a variable enables the effect of a variable on the outcome to differ between subjects. For example, a random effect could also be a random slope for a categorical variable. In general, in a mixed model, all of the variables of interest are added as fixed effects, but at least one and sometimes several of the fixed effects variables may also be added as random effects variables [19]. Therefore, the idea is that the values of a given random effect in the sample are a random sample of all possible values in the broader population (e.g., people in the sample are a random sample of people in the population). Moreover, in longitudinal studies, time or a time-varying covariate X is often an explanatory variable of interest, and the associations between explanatory variables and responses may vary between subjects. A model that allows heterogeneity in the intercept and heterogeneity in the magnitude of the slope between subjects is referred to as the random intercept and slope model. The random intercept and slope model is given by

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{i0} + b_{i1} t_{ij} + \varepsilon_{ij}$$

where t_{ij} is the time variable used as a predictor in the model.

A more general form of the mixed model is expressed as

$$Y_{ij} = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + b_{i0} + b_{i1} X_{i1} + \ldots + b_{ip} X_{ip} + \varepsilon_{ij}$$

where Y_{ij} is an outcome variable that indicates the j^{th} measurement on the j^{th} subject, $X_{ij}, j = 1,2,\ldots,p$ are the predictor variables, $\beta_0, \beta_1, \ldots, \beta_p$ are fixed effects, $b_{i1}, \ldots, b_{ip}$ are random effects, and ε_{ij}'s are residuals.

In the current model, the square root of CD4 count is used as the outcome because this transformation satisfies the normality assumption better than the untransformed CD4+ cell counts. Hence the model of interest is

$$\sqrt{CD4_{ij}} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1}) X_{1ij} + \ldots + (\beta_p + b_{ip}) X_{pij} + \varepsilon_{ij},$$

where

$\beta_0, \beta_1, \ldots, \beta_p$ are fixed effects, $b_{i1}, \ldots, b_{ip}$ are random effects, and ε_{ij}'s are residuals

The general matrix specification of the mixed model is

$$\underset{N\times 1}{Y} = \underset{N\times p}{X}\ \underset{p\times 1}{\beta} + \underset{N\times r}{Z}\ \underset{r\times 1}{U_i} + \underset{N\times 1}{\varepsilon}$$

with $i = 1, \ldots, n$ individuals and $j = 1, \ldots, N$ observations for individual i. Thus, Y is a $N \times 1$ vector of the response variable, $X = [X_{i1}, \ldots, X_{ip}]$ is $N \times p$ known design matrix that includes covariates for the fixed effects, β is $p \times 1$ vector of fixed effects parameters, $Z = [X_{i1}, \ldots, X_{ir}]$ is $N \times r$ known design matrix for random effects, U_i is $r \times 1$ vector of random effects from a normal distribution with variance-covariance matrix G, and ε is $N \times 1$ error vector from a normal distributionwith variance-covariance matrix R[19].

Assumption: U and are independent and each is normally distributed.

$$E\begin{bmatrix} U \\ \varepsilon \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ and } cov\begin{bmatrix} U \\ \varepsilon \end{bmatrix} = \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} U \\ \varepsilon \end{bmatrix} \sim \mathcal{N}\left[\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix}\right]$$

$$Y \sim N(X\beta, V = ZGZ' + R)$$

The distribution of Y is a multivariate normal distribution i.e. the vector of outcomes $Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$ is a multivariate normal distribution with mean vector $X\beta$and variance-covariance non-singular matrix V and its probability density function (pdf) is

$$f(Y) = (2\pi)^{-n/2} |V|^{-1/2} exp\left[\frac{-1}{2}(Y - X\beta)' V^{-1} (Y - X\beta)\right]$$

The log-likelihood of Y under this model is

$$l(\beta, V) = \frac{-n}{2}\log(2\pi) - \frac{1}{2} log|V| - \frac{1}{2}(Y - X\beta)' V^{-1} (Y - X\beta)$$

$$= \frac{-1}{2}\{n log(2\pi) - log|V| + (Y - X\beta)' V^{-1} (Y - X\beta)\}$$

Therefore, the maximum likelihood estimate (MLE) of $(\beta, V)$is the one that maximizes the right-side of the above expression [19].

Covariance or correlation structures that are most commonly used for longitudinal data analysis are compound symmetry (CS), unstructured (UN), First-order Autoregressive (AR (1)), and Toeplitz (Toep). These four common covariance structures are summarized in [5,7,8,16,19-22].

To decide which mixed-effects model fits the data best, we can use likelihood-based methods, i.e., either the likelihood ratio test (LRT) or Information Criteria (IC) such as Akaike Information Criteria (AIC) or Bayesian Information Criteria (BIC) method. The LRT, which is based on χ^2-distribution can be used to test nested models. The model with the lowest AIC and BIC is the best fitting model. That is, the AIC and BIC can be used to compare models such that the smaller of any of these, the better between two or more competing models. The IC method is more general to compare two or more competing non-nested models. However, the LRT is the best method to compare nested models [23].

In mixed-models, we use maximum likelihood (ML) to estimate the fixed effects, the standard errors of the fixed effects, and the variance of the random effects. The likelihood of mixed effect models can be time-consuming computationally, but with advances in

statistical software, this has become an easily manageable problem. Often the likelihood is solved by iteration until convergence. However, under ML estimation the residual variance and variance of random effects are underestimated thus instead the restricted maximum likelihood (REML) estimation gives unbiased estimates of variance parameters by taking into account the degrees of freedom used to estimate the fixed effects hence variance parameter estimates are generally larger than those from ML estimation. However, REML uses the covariate mean structure (the number of fixed effects) in the model estimation steps. That means we use REML when we are comparing two models that differ only in random effects (see page 352 in Der and Everitt, 2012) [4,24].

In general, when testing mixed-effects models that differ in variance components, we could either use REML or ML since they both give interpretable LRT and IC for such a comparison. However, testing and comparing models that differ in fixed effects, then only ML, will provide us with interpretable LRT and IC. However, ML does not take into account the degrees of freedom for the loss of fit in the estimation of parameters, but REML does [19,20].

Results

Data for this study were obtained from the CAPRISA 002: Acute infection Study, which was initiated between August 2004 and May 2005[13]. The baseline characteristics of the datasets are given in Table 1. From a total of 235 women, 105 (44.7%) were residing around Vulindlela (rural site), and 130 (55.3%) were residing around eThekwini (Durban, urban site), KwaZulu-Natal, South Africa. The average age at enrollment and baseline CD4+ cell counts was 27.15 years (range 18-59) with a standard deviation of 6.56 and 570 (range 182- 1575) with a standard deviation of 229.6, respectively. The average follow-up time was 2.69 years, and the majority of the women 182 (77.4%) had a stable partnership. Furthermore, from the total women included in the study, the majority of the 224 (95.3%) completed secondary/high education, and most of the women (78.8%) were self-reported sex workers[13,34]. There were a total of 7129 observations from the 235 women, which consists of a minimum of four and a maximum of sixty-one measurements of CD4+ cell counts, among the subjects which were measured at different time points indicating that the number of measurements over all subjects was not equal. Further apart from an unequal number of measurements across individuals, measurements were not taken at fixed time points, which implies the CAPRISA 002: Acute Infection Study is a highly unbalanced longitudinal data set that requires carefully designed modelling approaches.

Figure 1 (left panel) shows that CD4+ cell count distribution is right-skewed, indicating non-normality; thus, a square root transformation to CD4+ cell count was performed to normalize the data, Figure 1 (right panel) shows that the square root transformed data conforms quite well to the normal distribution.

Table 1: Baseline characteristics of the motivated data set (CAPRISA 002), 2004-2018.

Variable	Total	Variable	Total
Number of women	235	**Marital Status**	
Place of residence		No partner	43 (18.3%)
Rural	105 (44.7%)	Stable partner	182 (77.4%)
Urban	130 (55.3%)	Many partners	10 (4.3%)
Age at Seroconversion (Years)			
Mean (Std. Deviation)	27.15 (6.56)	**Educational Attainment**	
<20	21 (8.9%)	Primary schools (grade 0-7)	11 (4.7%)
20-29	150 (63.8%)	Secondary schools (grade 8-12)	224 (95.3%)
30-39	50 (21.3%)	**Baseline CD4+ cell counts (**cells/**µL)**	
40-49	12 (5.1%)	Mean (Std. Deviation)	570 (229.6)
≥ 50	2 (0.9%)	**Baseline HIV viral load (**cells/**µL)**	
Baseline Body Mass Index		Undetectable VL (< 50)	1 (0.4%)
Underweight	14 (6%)	Low VL (50<VL<10000)	74 (31.5%)
Normal weight	173 (73.6%)	Medium VL (10000<VL<100000)	94 (40%)
Overweight	41 (17.4%)	High VL (≥100000)	66 (28.1%)
Obese	7 (3%)		

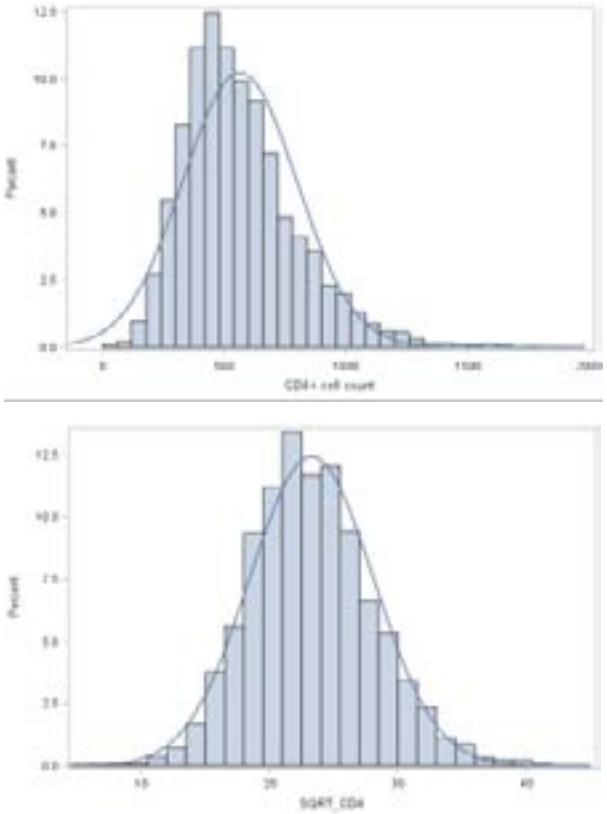


Figure 1: Distributional properties plot for original and square root transformed CD4 trajectories

The spaghetti plots in Figure 2 illustrate the actual CD4+ cell count measurements for randomly chosen participants over time across pre and post ART initiation groups. Since plots with all individual curves can be hard to distinguish for large sample size, we randomly chose 15 participants to construct such individual plots. From Figure 2, it can be seen that there is a decreasing trend of CD4+ cell count overtime on patients before Highly Active Antiretroviral Therapy (HAART) initiation, but an increasing trend of CD4+ cell count overtime for the same 15 randomly chosen patients initiated on HAART. Figure 2 also shows that there is evidence of variability between individuals as well as variability within individuals. Besides, the individual profiles are not all of the same lengths, an indication of incompleteness and missing data due to dropout or attrition.

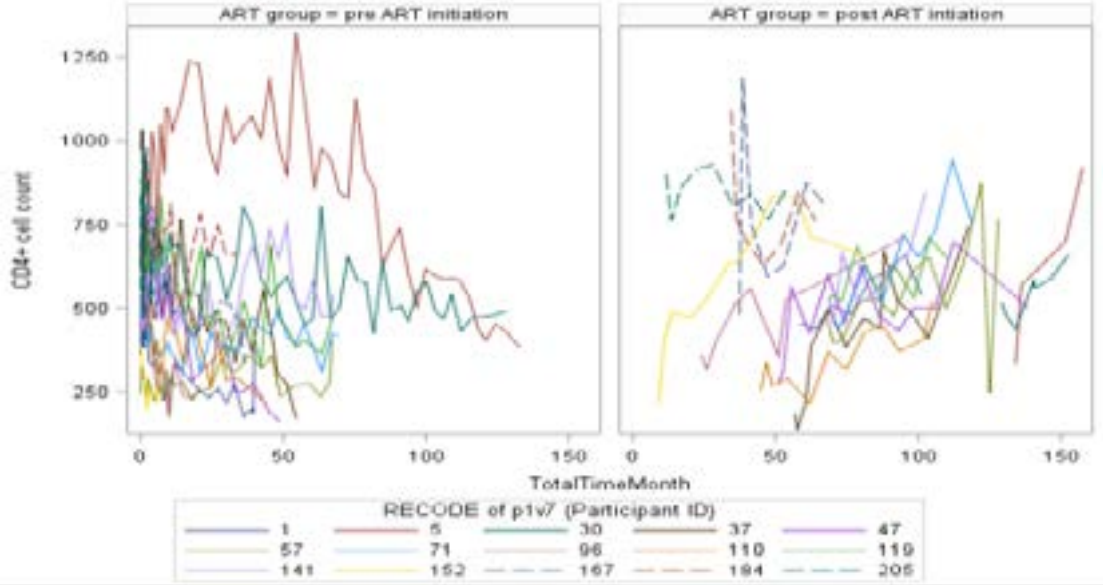


Figure 2: Individual profiles plot of CD4+ count for the same 15 randomly selected individuals before and after HAART.

Figure 3 shows an array of individual series from the CAPRISA 002: AI study. In each panel, the observed CD4 count for a single subject is plotted against the times that measurements were obtained. Such plots permit assessment of the person response patterns and whether there is substantial heterogeneity within the trajectories. Figure 3 shows that there can be variation in the "level" of CD4 count for subjects. Subject PID=5 in the first row second from left has CD4 counts greater than 500 for almost all times while PID= 110 in the third row lower-left corner has all measurements below 500. Moreover, PID=30 in the first row third from left has all measurements almost constant around 500. Further, individuals profile plots can be evaluated for the change over time [6]. Figure 3 shows that most subjects are either relatively stable in their measurements over time, or tend to be increasing.

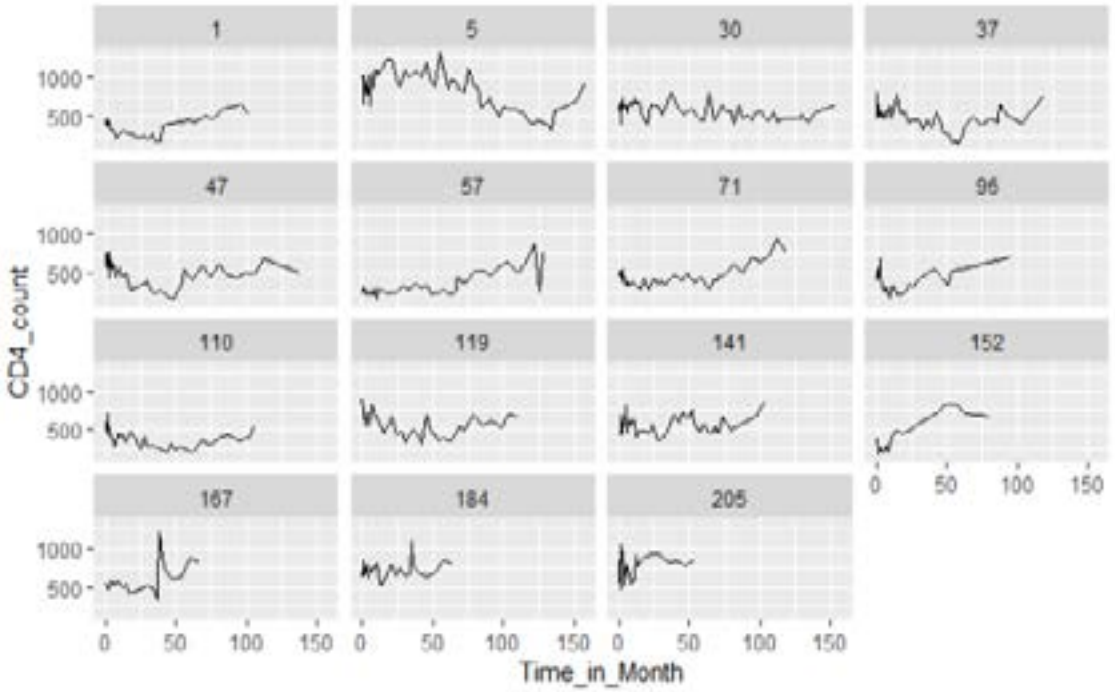


Figure 3: A sample of 15 individual CD4 trajectories versus time from the CAPRISA 002 AI Study

Figure 4 shows the mean CD4 trajectories overtime for the pre and post ART initiation groups in the CARISA 002: AI study. Overall the mean plots suggest that patients initiated on HAART have significant quadratic growth in the evolution of CD4 count over time as what we would expect. Furthermore, the plots exhibit non-linearity implying factors that control the nonlinear effect that may need to be incorporated in the model.

The inferential focus of this study is on the mean response of a square root transformation to CD4+ cell count measure. First, an appropriate selection of the random effects was also performed. That is the appraisal as to which of the nonlinear components (the intercept, time, or square root of time) ought to have a random component was made. To have a valid inference about the mean structure, the covariance structure must be incorporated into the statistical model [25]. Hence, following the selection of random components, a comparison of covariance structure was made in the study.

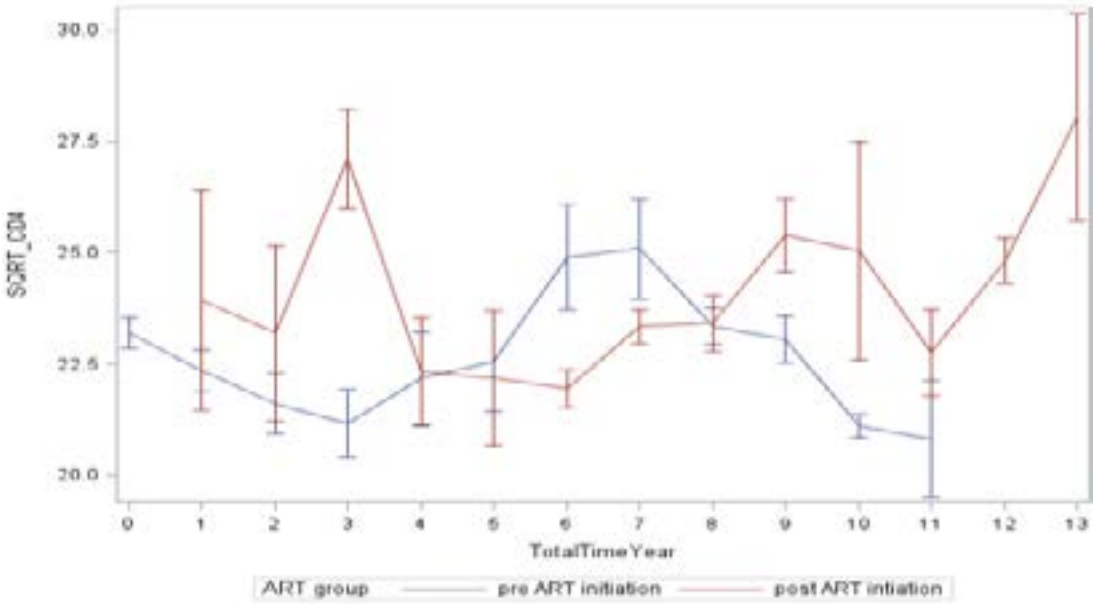


Figure 4: Mean CD4 trajectories over time by ART Initiation group, CAPRISA 002 AI study

The following random effect models, which have the same fixed effects, were fitted for testing:
Model 1: Intercept, Time, Square root of time *(Random intercept and slope model)*
Model 2: Time, Square root of time *(Random slope model)*
Model 3: Time only *(Random slope model without quadratic effect)*
Model 4: Intercept only *(Random intercept model)*
All models were fitted using the REML estimation procedure, and model comparison is made using different Information Criteria. The AIC statistics show that the random intercept and slope model is the preferable model among models listed above (Table 2).

Table 2: Model comparison using IC for random effects using REML estimation

Random effect models	Information Criteria						
	Params	-2log	AIC	AICC	HQIC	BIC	CAIC
Model 1	**4**	**34392.7**	**34400.7**	**34400.7**	**34406.3**	**34414.6**	**34418.6**
Model 2	3	36567.8	36573.8	36573.8	36577.9	36584.1	36587.1
Model 3	2	39832.4	39836.4	39836.4	39839.2	39843.3	39845.3
Model 4	2	36363.7	36367.7	36367.7	36370.5	36374.6	36376.6

To validate the random intercept and slope model (Model 1), a panel of conditional studentized residuals for the square root CD4+ count was used. The result is presented in Figure 5. The panel consists of a scatterplot of the residuals, a histogram with normal density, a Q-Q plot, and summary statistics for the residuals and the model fit. The residuals were randomly dispersed around zero, suggesting that their mean was approximately zero. The histogram follows a normal distribution indicating a constant variance. Hence, the fulfillment of the assumption that the error term ε_{ij} was normally distributed with mean 0 and variance σ^2.

Table 3 shows the comparisons between the four different covariance structures that were considered in the model using REML under the same fixed effects model. The Information Criteria was used to compare models for the structure that gives a better fit.

The estimated unstructured covariance parameter determines the matrix ($\hat{D}$) along with the estimated variance of the random error term (, respectively, are given below for Model 1:

$$\hat{D} = \begin{bmatrix} 20.1224 & 0.09786 & -2.4719 \\ 0.09786 & 0.01849 & -0.1705 \\ -2.4719 & -0.1705 & 1.9686 \end{bmatrix} \text{and} \hat{R} = var(\varepsilon_{ij}) = 5.7063$$

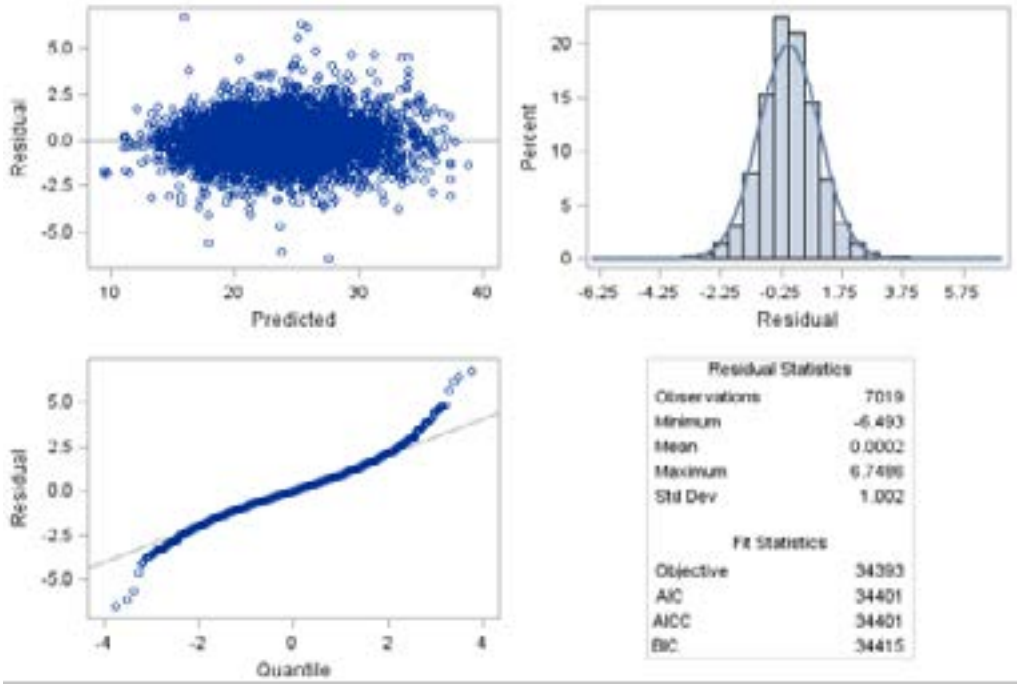


Figure 5: Panel of conditional studentized residuals for the square root of CD4 count

Table 3: Comparisons of covariance structure

Covariance Structure	Information Criteria						
	Params	-2log	AIC	AICC	HQIC	BIC	CAIC
AR(1)	3	35675.6	35681.6	35681.7	35685.8	35692.0	35695.0
CS	3	35671.5	35677.5	35677.5	35681.7	35687.9	35690.9
Toep	4	35671.4	35679.4	35679.4	35685.0	35693.2	35697.2
UN	**7**	**34087.1**	**34101.1**	**34101.1**	**34110.8**	**34125.3**	**34132.3**

Table 4 shows the REML estimates for the fixed effects of the random intercept and slope model (Model 1). Fitted conditional model or the subject-specific profile of the CD4+ count measure overtime 't' for the two ART initiation groups can be summarized as follows:

For post ART initiation group

$$\hat{Y}_{t'} = 26.7535 + 0.09015(time') - 0.9554(sqrt\_time')$$

For pre ART initiation group

$$\hat{Y}_{t'} = 24.3062 + 0.09015(time') - 0.9554(sqrt\_time')$$

Table 4: Fixed effect estimates of Model 1 for unstructured covariance structure

Effect	DF	Estimate	SE	Pr > |t|	95% C.I for Estimate
Intercept	234	24.3062	0.3055	<.0001	(23.7043, 24.9081)
Time in month	6781	0.09015	0.01072	<.0001	0.06913, 0.1112)
Sqrt_Time	6781	-0.9554	0.1036	<.0001	(-1.1586, -0.7523)
ART Initiation (Post)	195	2.4473	0.1348	<.0001	(2.1815, 2.7131)

The above fitted conditional models are extended to incorporate the impact of patient's age, educational status, number of sex partners, baseline BMI, baseline viral load, and ART initiation group with the square root of CD4 count as the response. In addition to this, two-way interaction effects were evaluated within the modelling process. But, none of the interaction effects was significant. The results of the effects of age, educa-

tional status, and the number of sex partners were not found to be significant. However, we incorporate them within the modelling process since factors with subject matter importance ought to be kept within the model to eliminate any confounding effects.

The results of the fixed effect estimates are presented in Table 5. As seen from Table 5, the model intercept $(\hat{\beta}_0)$is equal to 25.2439, which is an estimate of the mean square root CD4 count at baseline (i.e., month=0) subject to other effects with covariate values set to zero in the model. The Month effect $(\hat{\beta}_1) = 0.06377$is the slope or rate of change in the mean square root CD4 count per unit increase in the month among HIV infected patients with other covariate values set to zero. In other words, the time (month) effect shows a significant positive effect on the mean CD4 count with a rate of 0.06377 (p-value <0.0001) units per month. Hence square root CD4 count increases by 0.06377 for every month among patients, showing low progress of CD4 count over time. The effect of the square root of time (p-value < 0.0001) is also significant but appears to have an opposite effect on the square root CD4 count in a cohort of HIV infected patients enrolled in the CAPRISA 002 Acute Infection Study. The estimate for post-HAART initiation shows a highly significant positive effect with a mean square root CD4 count of 2.1104 units higher than the pre-HAART state. This implies, among patients in the post-HAART initiation group, their mean square root CD4 count increased by 2.1104, but this is not a slope. Relative to patients with normal weight status, patients with higher BMI (Obese) show a highly significant positive effect (p-value<0.0001) with 8.0201 square root CD4 count higher than the reference group (Table 5). However, underweight patients (patients with low BMI) show no significant effect compared to the reference grop. After the patients had been initiated on HAART, the average square root CD4 count among patients with a high value of the viral load at baseline is -3.2552 (p-value<0.0001) units lower compared to patients with low viral load at baseline. Moreover, after the patient had been initiated on HAART, the average square root CD4 count among patients with a medium viral load category at baseline is decreased by 1.5696 (p-value=0.0029) units compared to the average square root of CD4 count among patients with low viral load at baseline. Implying that patients with high and medium viral load at baseline have significantly lower mean CD4 count compared to patients with low viral load at baseline.

Table 5: Fixed effect estimates of the full Model

Covariates	Estimate	SE	Pr > |t|	95% C.I for Estimate
Intercept	25.2439	0.6040	<.0001	(24.0536, 26.4342)
Time in month	0.06377	0.009142	<.0001	(0.04585, 0.08169)
Sqrt_Time	-0.6674	0.09020	<.0001	(-0.8442, -0.4906)
ART Initiation (Post)	2.1104	0.1647	<.0001	(1.7855, 2.4353)
Baseline BMI category (ref.=Normal weight)				
Obese	8.0201	1.2896	<.0001	(5.4788, 10.5614)
Overweight	0.4966	0.5799	0.3927	(-0.6461, 1.6394)
Underweight	0.2486	0.9131	0.7856	(-1.5508, 2.0481)
Baseline HIV viral load category (ref.= Low VL)				
High VL	-3.2552	0.5633	<.0001	(-4.3652, -2.1452)
Medium VL	-1.5696	0.5211	0.0029	(-2.5965, -0.5426)
Undetectable	1.3418	3.3359	0.6879	(-5.2321, 7.9157)
Number of sex partner (ref.= Stable partner)				
Many partners	-1.4706	1.0859	0.1770	(-3.6105, 0.6693)
No partner	-0.6478	0.5791	0.2645	(-1.7889, 0.4933)
Age group (ref.= < 20)				
20-29	0.06144	0.4231	0.8847	(-0.7742, 0.8971)
30-39	0.1611	0.4780	0.7366	(-0.7831, 1.1053)
40-49	0.2491	0.6420	0.6985	(-1.0190, 1.5172)
50-59	-1.0100	1.0149	0.3212	(-3.0147, 0.9946)
≥ 60	-0.7631	1.9554	0.6969	(-4.6254, 3.0991)
Education attainment (ref.= Secondary or high school)				
Primary school	0.08077	1.0585	0.9392	(-2.0052, 2.1668)
Residence of participant (ref.= Urban)				
Rural	-0.2647	0.4539	0.5604	(-1.1593, 0.6298)

Spatial covariance structure measures the actual distance or variation among observations in space that are identified as unequally spaced longitudinal data [16,26]. The objective of including spatial covariance structure in mixed-effects models is to account for spatial variability (heterogeneity), failure to do so can result in erroneous conclusions. The spatial covariance structure model is $C(h) = C_0 + \sigma^2\rho(h)$

Where C_0, σ^2, and $\rho(h)$ indicates the *nugget*, the sill and the range (covariance structure model), respectively [16,26]. Table 6 shows a comparison of the three commonly used spatial covariance structures: spatial exponential structure (SP(EXP)), spatial spherical structure (SP(SPH)), and spatial Gaussian structure SP(GAU). Since the exponential model has the smallest information criteria statistics and the smallest -2log ℓ suggests that the SP(EXP) structure is the best of the three spatial covariance models (Table 6).

Table 6: Comparison of spatial covariance models

Spatial covariance	Model Fitting Criteria						
	Params	-2log	AIC	AICC	HQIC	BIC	CAIC
SP(EXP)	9	33024.5	33042.5	33042.6	33055.1	33073.6	33082.6
SP(SPH)	9	33039.1	33057.1	33057.1	33069.6	33088.2	33097.2
SP(GAU)	9	33162.1	33180.1	33180.1	33192.7	33211.2	33220.2

The estimate of the *sill* (σ^2) is 9.7063, reported as "Variance", which corresponds to the variance of observation (Table 7). The estimated *range* ($\rho(h)$) is 31.1376, which appears as "SP(EXP)", which is the practical range or distance at which the spatial autocorrelation in the exponential model is three times this amount, $3 \times 31.1376 = 93.4128$. That is, observations separated by more than 93.4128 distance units are not spatially correlated. In other words, the distance units indicate that observations within a participant that are close in time to be more correlated than observations farther apart in time. The estimated *nugget* (C_0) is 3.4986, which appears as "Residual", that is the value at which $h = 0$ or defined as *Intercept* in the spatial covariance structure model.

Table 7: Covariance Parameter Estimates of the full model

Cov Parm	Estimate	SE	Z Value	Pr>Z
UN(1,1)	3.3317	2.6772	1.24	0.1067
UN(2,1)	0.05870	0.04370	1.34	0.1792
UN(2,2)	0.004944	0.001733	2.85	0.0022
UN(3,1)	-0.3405	0.4031	-0.84	0.3983
UN(3,2)	-0.05410	0.01654	-3.27	0.0011
UN(3,3)	0.6223	0.1798	3.46	0.0003
Variance	9.7063	2.3528	4.13	<.0001
SP(EXP)	31.1376	9.4724	3.29	0.0005
Residual	3.4986	0.1008	34.70	<.0001

Figure 6 indicates the predicted profile plot for the average number of CD4+ cell, based on Table 5 results obtained by the fitted mixed-effects model. The predicted values closely matched the observed CD4+ count mean profile, with an R2=0.75, suggested that the overall model fit was good (Figure 6).

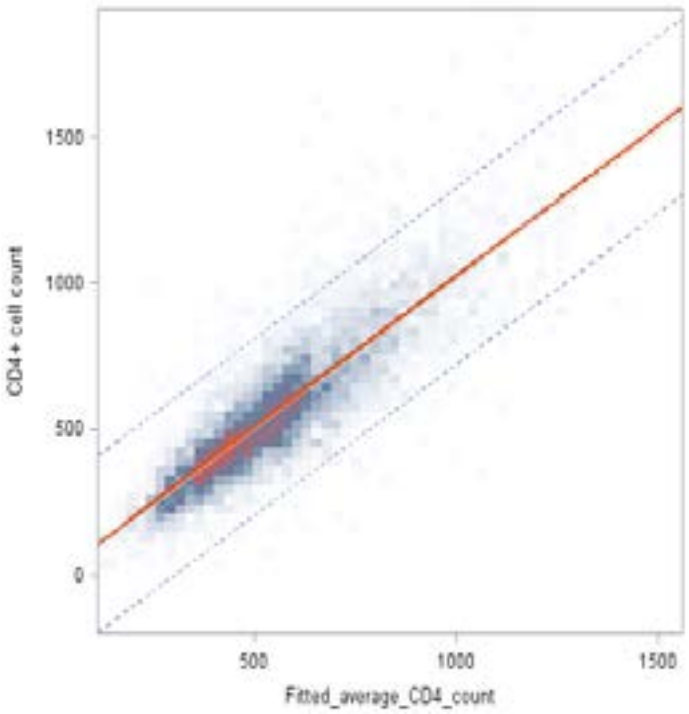


Figure 6: Heat map of fitted average by observed CD4 count overlaid with the fitted line

The fitted solid line in Figure 6 also indicates the estimated regression line between the observed CD4+ count and fitted values (Fitted= 148.07+0.7259 observed), and the two dashed lines show both 95% confidence interval and prediction interval.

The overall influence diagnostic and diagnostics for the fixed effects are displayed graphically hereunder in Figure 7-11. Figure 7 shows the needle plot of the Restricted Likelihood Distance (RLD) for the response variable (square root of CD4+ count). The RLD plot suggests that the overall influence of patients 5, 12, 29, 32, 55, 84, and 188 stands out compared to those of the rest of the patients (Figure 7).

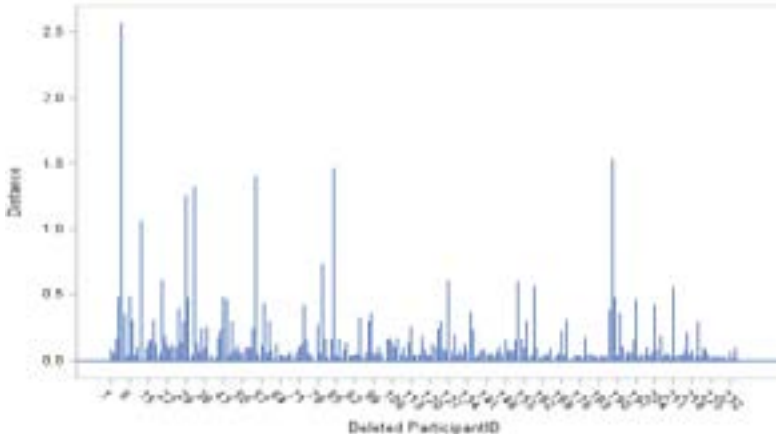


Figure 7: Restricted Likelihood Distance

PRESS statistics are sums of squared PRESS residuals in the deletion sets (Schabenberger, 2005). Figure 8 shows the scatter plot of the PRESS statistics for the square root of the CD4+ count. Large values of the PRESS statistic for patients 5, 60, 84, 127, and 189 are noted.

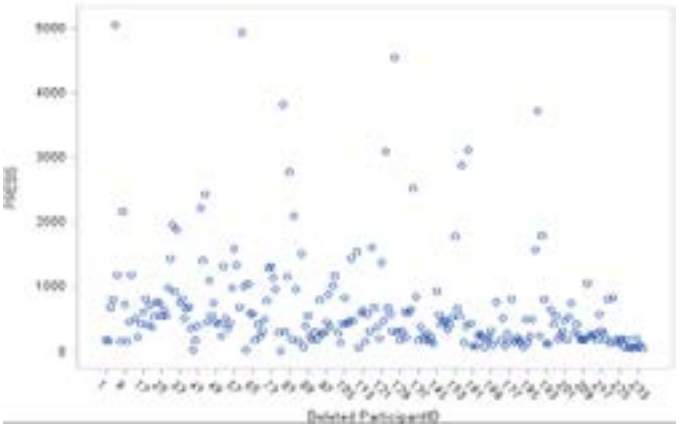


Figure 8: PRESS Statistics

A panel of influence statistics for fixed effects and covariance parameters is presented in Figure 9. Cook's D statistics measure the influence on the vector of parameter estimates and the CovRatio statistic measures influence on the covariance matrix of the parameter estimates. The patients with the most substantial effect on the fixed effect estimates are 5, 32, and 55 (Cook's D Fixed effects). Cook's D Covariance parameters indicate that the influence of patient 12, 84, and 188 far exceeds those of other subjects in the study data sets. This is expected since their RLD is substantial, while their impact on the fixed effects was rather moderate. The CovRatio Covariance Parameters also shows that in the absence of those patient's observations, especially patient 84 and 188, the covariance parameters may be estimated much more precisely. Note that there are other sets of observations, besides those patients listed above, that exerts influence on the chosen model (Model 1).

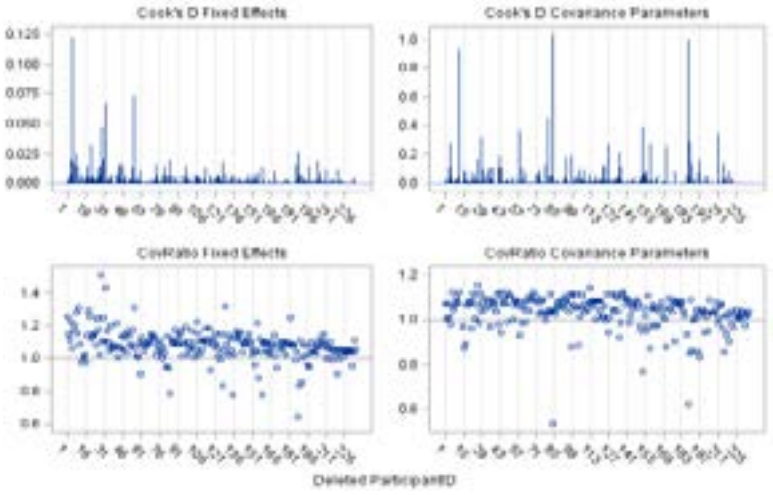


Figure 9: Influence statistics for the square root of CD4+ count

A panel of deletion estimates for the response variable is displayed in Figures 10 and 11 to examine how the individual parameter estimates and covariance parameters, respectively, react to the removal of the influential sets of observations[27]. Each cell in the panel (Figure 10) displays the estimates of few fixed effects that were included in the fitted model and each cell in Figure 11 displays estimates of the 3x3 variance-covariance matrix of the random coefficients and the estimate of SP(EXP) parameter following removal of sets of influential observations. Reference lines are drawn at the complete-data parameter estimates.

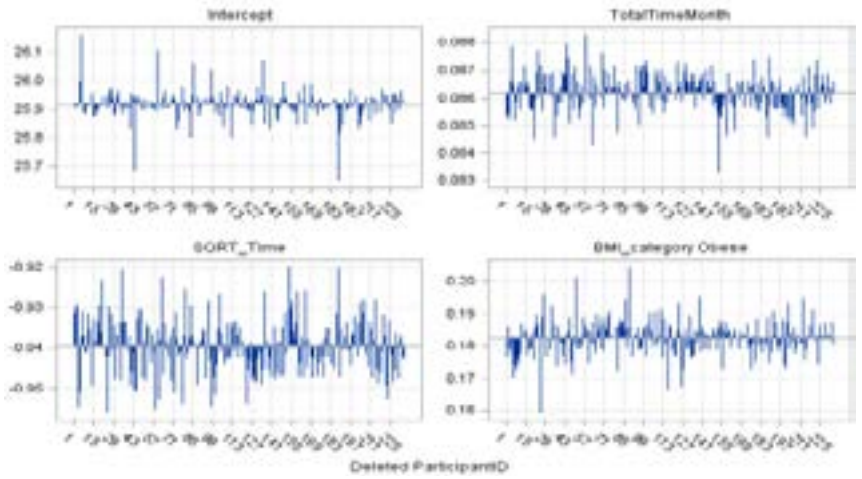


Figure 10: Fixed effects deletion estimates for square root of CD4+ count

The focus of Figure 10 is on the behavior of individual parameter estimates that react to the removal of influential cases. Specifically, subjects 5, 44, 60, and 188 indicate a substantial impact on the model fit of the intercept. However, the removal of these subjects does not at all influence the displayed fixed effects. On the other hand, subject 27 is identified as an additional influential case since it has a strong impact on the Obese BMI category (Figure 10). Subjects 5, 29, 73, and 85 are also identified as influential cases since their presence in the data reduces the estimate of SP(EXP) parameter (Figure 11), substantially reducing the degree of correlation among data points from any patient. On the other hand, observation from subject 12 has the opposite effect. The temporal correlation drops when the impact of this patient's data is removed.

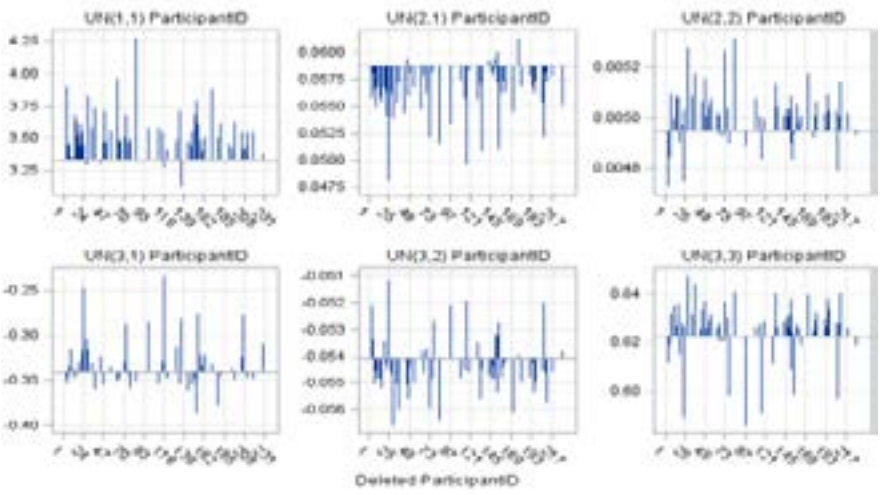


Figure 11: Covariance parameter deletion estimates for square root of CD4+ count

Discussion and Conclusion

Mixed-models are one of the special statistical models that are useful in understanding longitudinal or repeated measures data. The models permit the examination of the changes over time within and between subjects. In the presence of fixed effects and random effects, the selection of an appropriate mixed model is more complicated than for a linear regression model. The fixed effect and the random effect structure are subordinate to each other, and the determination of one influences the other[28]. In this study, a step-up model selection procedure was applied to find a reasonable model that fits the data, primarily since this procedure begins with the simplest possible model and is built up by including more covariates within the model and hence does not have much numerical issue [1,18,28]. In this study, the model where the intercepts and slopes were considered as random effects consolidated with the UN covariance structure was used. The results show that the prognosis of the CD4 count of a patient is significantly increased after the patient had been initiated on HAART as what we would anticipate. The impact of HIV-infected pa-

tients with the predominance of obese nutrition status (higher BMI) at baseline showed significance after patients had been initiated on HAART. Therefore, we ought to pay more consideration to the BMI of HIV-infected patients before and after HAART initiation. This may inform future techniques in studying the progression and the immunologic responses to treatment, but that does not infer that patients with higher BMI ought to be clinically ignored. Instead, based on this study and other findings, it appears that BMI contributes to some degree to drug metabolism and consequently influencing the proficiency of HAART[29,30]. Moreover, our results also showed that the impact of patients with higher viral load before the patient had been initiated on HAART significantly reduced their CD4 count. Therefore, effective HAART initiation after HIV exposure is necessary to suppress the increase of viral loads to induce potential ART benefits that accrue over time.

The results of the influence diagnostics analysis for the CAPRISA 002 Acute Infection study using the chosen mixed-effects model was also performed. Several cases were identified as influencing the analysis of the fitted model. Influence diagnostics analysis is essential for statistical analysis to determine how individual observations or sets of observations are influential that their presence or absence from the data impacts the analysis [31]. The goal of influence analysis is not to determine observations for removal from the analysis, but to determine which cases exert undue influence on the analysis. Eliminating certain subjects from the data and base the final analysis on only the remainder is usually not the right action to take. The results of a diagnostic influence analysis can be seen only in light of the model we are working with [16].

Moreover, the data showed evidence of strong individual-specific effects on the evolution of CD4+ counts. The diagnostic plots also suggested a significant individual heterogeneity between subjects both before and after HAART initiation. Thus this may suggest that prescribing a common treatment or intervention overall patients may not be the best strategy. More research may be required to understand what factors cause patients to respond differently to treatment intervention, and such information may help to design treatment and intervention strategies that may be more efficient to a specific group of patients rather than one treatment/intervention fits all strategy.

The models depicted in this study may empower the description of the effect of several covariates on the square root CD4 count of HIV-infected patients utilizing all accessible information. We believe that this sort of analysis can be valuable to address several important issues in public health as well as offer assistance in observing patients and checking the viability of their medications. In this study, we have concentrated on the transformed normalized response data, which is the square root of CD4 count, that is continuous and conditional on the explanatory variables, and random effects have a normal distribution. Mixed models with random effects can also be applied to non-normal responses.

Abbreviations

CAPRISA: Centre of the AIDS Programme of Research in South Africa; AI: Acute Infection; HIV: Human Immunodeficiency Virus; AIDS: Acquired Immune Deficiency Syndrome; CD4: Cluster of Difference 4 cell (T-lymphocyte cell); VL: Viral Load refers to the number of HIV copies in a milliliter of blood (copies/ml); ART: Antiretroviral Therapy; ARV: Antiretroviral (drug); HAART: Highly Active Antiretroviral Therapy; WHO: World Health Organization; UNAIDS: Joint United Nations Programme on HIV/AIDS; REML: Restricted Maximum Likelihood; UN: Unstructured covariance structure; MCAR: Missing Completely at Random; BMI: Body Mass Index; IC: Information Criterion.

Acknowledgments

We gratefully acknowledge CAPRISA for giving us access to the CAPRISA 002: Acute Infection Study data. CAPRISA is funded by the National Institute of Allergy and Infectious Diseases (NIAID), National Insttutes for Health (NIH), and U.S. Department of Health and Human Services (grant: AI51794). The authors would also like to thank Dr. Nonhlanhla Yende Zuma (Head of Biostatistics unit at CAPRISA) for her cooperation, assistance, and technical support.

Financial support

This work was supported through the DELTAS Africa Initiative and the University of KwaZulu-Natal. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [grant 107754/Z/15/Z], DELTAS Africa Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB) programme] and the UK government.

Disclaimer
The views expressed in this publication are those of the author(s) and not necessarily those of CAPRISA, AAS, NEPAD Agency, Wellcome Trust, or the UK government.

Authors' contributions
AAY acquired the data, performed the analysis, and drafted the manuscript. AAY, SFM, HGM, and DGA designed the research problem. All authors discussed the results and implications and commented on the manuscript at all stages. All authors contributed extensively to the work presented in this paper. All authors read and approved the final manuscript.

Ethics approval
Ethical approval for the CAPRISA 002: Acute Infection Study was obtained from the Research Ethics Committee of the University of KwaZulu-Natal (E013/04), the University of the Witwatersrand (MM040202) and the University of Cape Town (025/2004). All participants provided written, informed consent to enroll inthe study.

Consent for publication
Not applicable.

Availability of data
The data used for this study can be obtained by requesting CAPRISA.

Competing interests
The authors declare that they have no competing interests, financial or otherwise.

References

1. Diggle P, Diggle PJ, Heagerty P, Liang K-Y, Heagerty PJ, Zeger S. Analysis of longitudinal data: Oxford University Press; 2002.
2. Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G. Longitudinal data analysis: CRC press; 2008.
3. Hox JJ, Moerbeek M, Van de Schoot R. Multilevel analysis: Techniques and applications: Routledge; 2017.
4. Der G, Everitt BS. Applied medical statistics using SAS: Chapman and Hall/CRC; 2012.
5. Hedeker D, Gibbons RD. Longitudinal data analysis: John Wiley & Sons; 2006.
6. Twisk JW. Applied longitudinal data analysis for epidemiology: a practical guide: cambridge university press; 2013.
7. Kincaid C, editor Guidelines for selecting the covariance structure in mixed model analysis. Proceedings of the thirtieth annual SAS users group international conference; 2005: SAS Institute Inc Cary NC.
8. Kowalchuk RK, Keselman H, Algina J, Wolfinger RD. The analysis of repeated measurements with mixed-model adjusted F tests. *Educational and Psychological Measurement*. 2004;64(2):224-42.
9. Liu X. Methods and applications of longitudinal data analysis: Elsevier; 2015.
10. Brown H, Prescott R. Applied mixed models in medicine. West Sussex, United Kingdom: John Wiley & Sons; 2014.
11. Taris T. A Primer in Longitudinal Data Analysis Sage. London; 2000.
12. Garrett N, Norman E, Leask K, Naicker N, Asari V, Majola N, et al. Acceptability of early antiretroviral therapy among South African women. *AIDS and Behavior*. 2018;22(3):1018-24.
13. Mlisana K, Werner L, Garrett NJ, McKinnon LR, van Loggerenberg F, Passmore J-AS, et al. Rapid disease progression in HIV-1 subtype C–infected South African Women. *Clinical Infectious Diseases*. 2014;59(9):1322-31.
14. Moosa Y, Tanko RF, Ramsuran V, Singh R, Madzivhandila M, Yende-Zuma N, et al. Case report: mechanisms of HIV elite control in two African women. *BMC Infectious Diseases*. 2018;18(1):1-7.
15. Duchateau L, Janssen P, Rowlands J. Linear mixed models. An introduction with applications in veterinary research: ILRI (aka ILCA and ILRAD); 1998.
16. Littell RC, Milliken GA, Stroup WW, Wolfinger RD, Oliver S. SAS for mixed models: SAS publishing; 2006.
17. Verbeke G, Molenberghs G. Linear mixed models for longitudinal data: Springer Science & Business Media.; 2009.
18. West BT, K. B. Welch, A. T. Galecki. Linear mixed models: a practical guide using statistical software: Chapman and Hall/CRC; 2014.
19. Rawlings JO, Pantula SG, Dickey DA. Applied regression analysis: a research tool: Springer Science & Business Media; 2001.
20. Hofer A. Variance component estimation in animal breeding: a review. *Journal of Animal Breeding and Genetics*. 1998;115(1-6):247-65.
21. Searle SR, Casella G, McCulloch CE. Variance components: John Wiley & Sons; 2009.
22. Wolfinger RD. Heterogeneous variance: covariance structures for repeated measures. Journal of agricultural, biological, and environmental statistics. 1996:205-30.
23. Loy A, Hofmann H, Cook D. Model choice and

diagnostics for linear mixed-effects models using statistics on street corners. *Journal of Computational and Graphical Statistics*. 2017;26(3):478-92.
24. Longford NT. Random coefficient models. *Handbook of statistical modeling for the social and behavioral sciences: Springer*; 1995. p. 519-70.
25. Melesse SF, Zewotir T. Modelling the effect of tree age and climatic factors on the stem radial growth of juvenile eucalypt clones. Bulletin of the Transilvania University of Brasov Forestry, Wood Industry, Agricultural Food Engineering Series II. 2017;10(1).
26. Zimmerman DL, Harville DA. A random field approach to the analysis of field-plot experiments and other spatial experiments. *Biometrics*. 1991:223-39.
27. Schabenberger O, editor Mixed model influence diagnostics. SUGI; 2005: Citeseer.
28. Melesse SF, Zewotir T. Additive mixed models to study the effect of tree age and climatic factors on stem radial growth of Eucalyptus trees. *Journal of Forestry* Research. 2020;31(2):463-73.
29. Li X, Ding H, Geng W, Liu J, Jiang Y, Xu J, et al. Predictive effects of body mass index on immune reconstitution among HIV-infected HAART users in China. *BMC Infectious Diseases*. 2019;19(1):373.
30. Palermo B, Bosch RJ, Bennett K, Jacobson JM. Body mass index and CD4+ T-lymphocyte recovery in HIV-infected men with viral suppression on antiretroviral therapy. *HIV Clinical Trials*. 2011;12(4):222-7.
31. Zewotir T. Multiple cases deletion diagnostics for linear mixed models. Communications in Statistics: Theory and Methods 2008;37(7):1071-84.
32. Whelan D. Gender and HIV/AIDS: taking stock of research and programmes. UNAIDS; 1999.
33. WHO, USAIDS, Unicef. Epidemiological Fact Sheet on HIV and AIDS Core data on epidemiology and response. South Africa; 2008.
34. van Loggerenberg, F. KM, C. Williamson, S. C. Auld, L. Morris, C. M. Gray, et al. Establishing a cohort at high risk of HIV infection in South Africa: challenges and experiences of the CAPRISA 002 acute infection study. *PloS One*. 2008;3(4):e1954.

SCIENTIFIC REPORTS
nature research

OPEN

Negative binomial mixed models for analyzing longitudinal CD4 count data

Ashenafi A. Yirga[1]✉, Sileshi F. Melesse[1], Henry G. Mwambi[1] & Dawit G. Ayele[2]

It is of great interest for a biomedical analyst or an investigator to correctly model the CD4 cell count or disease biomarkers of a patient in the presence of covariates or factors determining the disease progression over time. The Poisson mixed-effects models (PMM) can be an appropriate choice for repeated count data. However, this model is not realistic because of the restriction that the mean and variance are equal. Therefore, the PMM is replaced by the negative binomial mixed-effects model (NBMM). The later model effectively manages the over-dispersion of the longitudinal data. We evaluate and compare the proposed models and their application to the number of CD4 cells of HIV-Infected patients recruited in the CAPRISA 002 Acute Infection Study. The results display that the NBMM has appropriate properties and outperforms the PMM in terms of handling over-dispersion of the data. Multiple imputation techniques are also used to handle missing values in the dataset to get valid inferences for parameter estimates. In addition, the results imply that the effect of baseline BMI, HAART initiation, baseline viral load, and the number of sexual partners were significantly associated with the patient's CD4 count in both fitted models. Comparison, discussion, and conclusion of the results of the fitted models complete the study.

Abbreviations

AI	Acute Infection
AIDS	Acquired immune deficiency syndrome
ART	Antiretroviral therapy
ARV	Antiretroviral (drug)
CAPRISA	Centre of the AIDS Programme of Research in South Africa
CD4	Cluster of difference 4 cell (T-lymphocyte cell)
GLM	Generalized linear model
GLMM	Generalized linear mixed model
HAART	Highly active antiretroviral therapy
HIV	Human immunodeficiency virus
MI	Multiple imputations
NBMM	Negative binomial mixed-effects model;
PMM	Poisson mixed-effects model
SE	Standard error
STD	Sexually transmitted disease
VL	Viral load refers to the number of HIV copies in a milliliter of blood (copies/ml)

After it is identified by scientists as the human immunodeficiency virus (HIV) and the cause of acquired immunodeficiency syndrome (AIDS) in 1983, HIV has spread persistently, triggering one of the most severe pandemics ever documented in human history. More than 75 million individuals have been infected with HIV, more than 32 million individuals have perished due to AIDS-related causes since the pandemic started, and 7000 new infections are reported daily. Worldwide, 37.9 million [32.7–44.0 million] individuals were HIV positive at the end of 2018. Approximately 0.8% [0.6–0.9%] of grownup persons in the age range fifteen to forty-nine years worldwide are living with HIV, even though the problem of the epidemic continues to vary sizably between nations and

[1]School of Mathematics, Statistics, and Computer Science, University of KwaZulu-Natal, Private Bag X01, Scottsville, Pietermaritzburg 3209, South Africa. [2]Institute of Human Virology, School of Medicine, University of Maryland, Baltimore, MD, USA. ✉email: ashu3argaw@gmail.com

regions[1]. Despite recent progressions in HIV prevention, care, and treatment, which has modestly decreased the total number of new infections and deaths every year, AIDS and AIDS-related illnesses are still among the driving causes of loss of life globally. Sub-Saharan Africa and Southern Africa, in specific, is right now the region most influenced by HIV/AIDS in the world[2]. The HIV crisis in South Africa is critical. Since South Africa is at the epicenter of the HIV/AIDS epidemic, South African concerns are worldwide concerns, and lessons learned in South Africa are lessons for the universal community.

HIV/AIDS and other STD have an obliterating effect on women's health, especially the well-being of younger ladies. "The consequences of HIV/AIDS attain beyond women's health to their part as mothers and caregivers and their commitment to the economic support of their families. The social, development, and health consequences of HIV/AIDS and other sexually transmitted illnesses should be seen from a gender perspective"[3–5]. "It needs to be emphasized that, except for sex-specific issues, treatment algorithms for HIV-Infected women do not differ from men's. Dialogs about the changing epidemiology of HIV will provide the clinician a system to decide who may be at high risk and to clarify the application of rules to avoid sequential HIV transmission. Even though antiretroviral recommendations presently remain the same for men and women, the survey of discoveries for early HIV infection and the individual difference in CD4 cell count/viral load of HIV-infected patient will permit the clinician to interpret prospective information appropriately and to address deception or distortion of this information by patients"[6–8].

"CD4 cell counts deliver a sign of the wellbeing of an individual immune system (body's natural defense system against pathogens, infections, and illnesses). It also provides information about disease progression. CD4 cells are white blood cells (in a cubic millimeter of blood) that play an essential role in the immune system. A higher number shows a stronger immune system. The CD4 cell counts of a person who does not have HIV can be anything between 500 and 1500. Individuals living with HIV who have a CD4 count over 500 are usually in good health. Individuals living with HIV who have a CD4 cell count below 200 are at high risk of developing serious illnesses[9]. HIV treatment is prescribed for all individuals living with HIV. It is particularly critical for patients with low CD4 count, which is superior to start treatment sooner, rather than later"[6]. The study of HIV infection at the acute stage is essential to the plan and advancement of HIV antibodies and techniques to attain an undetectable level of the infection without ART or a functional remedy. Researchers have managed to find out about the early events following infection by diagnosing HIV within a month, weeks, or even days of infection. Moreover, humans dwelling with HIV who are not on treatment or who are not virally suppressed can also have a compromised immune system (measured by a low CD4 count) that makes them at risk of the new and ongoing coronavirus disease 2019 (COVID-19) pandemic, opportunistic infections, and underlying illnesses. Whereas analysts accept that early diagnosis and prompt treatment of HIV are the stepping stones to a functional remedy, more studies are required to understand better the adaptive, innate, and host responses that regulate viral load set-point and subsequently diagnosis and infectiousness.

Count data are ubiquitous in public health investigations. This sort of data assumes only positive integer values (i.e., 0, 1, 2, …). The most commonly used method for count data is the Poisson distribution and its related enhancement, such as the Poisson-gamma mixture, which considers over-dispersion and heterogeneity in the model. This paper's main contribution is the inclusion of the links between CD4 cell count and influencing covariates of biometric and demographic factors. Therefore, this study aims to cope with the statistical challenges of over-dispersion and incorporate within-subject correlation structures by applying NBMMs to longitudinal CD4 count data from the CAPRISA 002 AI Study and also detecting factors that are significantly associated with the response variable.

Materials and methods

Data description. This study makes use of data from the CAPRISA 002 AI Study. The study was conducted on HIV-infected women at the Doris Duke Medical Research Institute (DDMRI) at the Nelson R Mandela School of Medicine of the University of KwaZulu-Natal in Durban, South Africa. Between August 2004 and May 2005, CAPRISA introduced a cohort study recurring high-risk HIV negative women to a follow-up study. In the case of the data used in this paper as part of an ongoing study, women infected with HIV are enrolled in the study early, followed intensely, and monitored carefully to examine disease progression and CD4 count/viral load evolution. One can refer to studies by Van Loggerenberg et al.[10] and Mlisana et al.[11] for details on the design, development, and procedures of the study population.

Methods. A linear model consists of a response variable Y, which is assumed to be normally distributed, and several predictors $(x_1, x_2, \ldots, x_p)$. Multiple regression analysis studies the linear relationships among two or multiple independent variables and one dependent (response) variable. The multiple regression model is given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i = \beta_0 + \boldsymbol{x_i'\beta} + \varepsilon_i = \beta_0 + \boldsymbol{\beta' x_i} + \varepsilon_i, i = 1, \ldots, n.$$

where y_i is the response variable, $\boldsymbol{x_i}$ is a $p \times 1$ vector of explanatory variables, β_0 is the intercept, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression coefficients, and $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, which is a random error of observation i. We can extend these multiple linear regression model ideas to generalized linear models (GLM) where the distribution of the outcome variable can include distributions other than normal. The outcome y_i can be continuous, dichotomous, count, ordinal, categorical, and so on as long as its distribution is from the exponential family. The exponential family of distributions incorporates numerous distributions that are valuable for viable modeling such as Poisson and Negative Binomial for count data; Binomial, Bernoulli, and Geometric for discrete data;

Gamma, Normal, Inverse Gaussian, Beta, and Exponential for the study of continuous response data set. More details on exponential family and related topics can be found in Dobson et al.[12].

A Poisson process is mainly used as an initial point for modeling the stochastic difference of count data around a theoretical expectation. However, in reality, the patient's data have more differences than using the Poisson distribution. The model's over-dispersion is accounted for because of different model assumptions about the variance changes with the expectation. To the value of statistical inferences, the choice of these assumptions has major consequences. Therefore, the negative binomial distribution parameterization is proposed because the method introduces various quadratic mean–variance relationships, incorporating the ones assumed in the most commonly used approaches.

The Poisson regression is a commonly-used statistical model for n responses $y_1, \ldots, y_n$ whose domain is non-negative integer values. Each y_i is modeled as an independent Poisson (λ_i) random variable and distributed as $y_i \overset{iid}{\sim}$ Poisson (λ_i), where the parameter λ_i controls the count rate in the ith outcome. Thus, a model for the Poisson rate parameter λ_i is given by

$$\ln \lambda_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

or equivalently,

$$\lambda_i = e^{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}} = e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}}$$

where $x_{i1}, \ldots, x_{ip}$ are a set of p explanatory variables, and $\beta = (\beta_0, \ldots, \beta_p)$ are the regression coefficients. The probability mass function (pmf) of the Poisson random variable with parameter λ_i is given by

$$f(y_i, \lambda_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}, y_i = 0, 1, 2, \cdots \tag{1}$$

Since $y_i \overset{iid}{\sim}$ Poisson (λ_i), as a consequence, the likelihood function is equal to the product of their pmf and the log-likelihood function can be derived by taking the natural logarithm of the likelihood function, become

$$= \sum_{i=1}^{n} \left[y_i \ln(\lambda_i) - \lambda_i - \ln y_i! \right]$$

where λ_i is defined in terms of $\beta_0, \ldots, \beta_p$ and the covariates $x_{i1}, \ldots, x_{ip}$ in Eq. (1), the log-likelihood function can be expressed as

$$\begin{aligned} \ell(\beta_0, \ldots, \beta_p) &= \sum_{i=1}^{n} \left[y_i \left(\sum_{j=0}^{p} \beta_j x_{ij} \right) - e^{\sum_{j=0}^{p} \beta_j x_{ij}} - \ln y_i! \right] \\ &= \sum_{i=1}^{n} \left\{ y_i \boldsymbol{x_i'} \boldsymbol{\beta} - \exp\left(\boldsymbol{x_i'} \boldsymbol{\beta} \right) - \ln y_i! \right\}. \end{aligned}$$

For a presentation of efficient computational methods for maximizing $\hat{\boldsymbol{\beta}}$, and $V\left[\hat{\boldsymbol{\beta}}\right]$, see Hilbe[13].

Suppose the response variable y_i follows a Poisson distribution with mean λ_i and there is no over- or under-dispersion, then $var(y_i) = \lambda_i$ that is the mean and variance are equal. The restriction (mean = variance) may not be satisfied with many real-world data. Sometimes the variance is greater than the mean, and this phenomenon is called over-dispersion. One such model that works in such a condition is the negative binomial regression model.

If there is over-dispersion $var(y_i) = \Phi\lambda_i$ and $\Phi > 1$. While if there is under-dispersion $var(y_i) = \Phi\lambda_i$ and $\Phi < 1$ that is $var(y_i) > E(y_i)$, in this case, the Poisson distribution is no longer suitable. The method of moments solution for the dispersion parameter Φ is found from the sample relation that is $var(y_i) = \hat{\Phi}\bar{y}$. Therefore, $\hat{\Phi} = \frac{\text{var}(y_i)}{\bar{y}}$, and then if $\hat{\Phi} > 1$, evidence of over-dispersion. Data may be over-dispersed if the Pearson Chi-Square (χ^2)/DF value is greater than 1.0. In general, when the value is greater than 2.0, it is an indication of over-dispersion, it requires remedial action[13,14]. Over-dispersed data can lead to underestimated SEs and inflated test statistics[13–16]. In such circumstances, the negative binomial model can be utilized, and therefore the formulation can be expressed as $y_i \sim NB(\mu_i, \mu_i[1 + \alpha\mu_i])$, where $\alpha(\alpha > 0)$ can be utilized to add flexibility, and plays the role of the scale parameter, for variance independently of the mean. The negative binomial model is a generalization of the Poisson model, which relaxes the restrictive assumption that the variance and mean are equal[13–15]. Just like the Poisson model, the negative binomial model is commonly utilized as a distribution for count data; however, it allows a variance higher than its mean. The most contrast between the NB and Poisson models is the extra parameter (scale parameter) that controls for the over-dispersion and, thus, the determination of the likelihood functions related to them[13,14]. Estimation of the parameters can be accomplished through likelihood maximization by employing a nonlinear optimization method[13,14]. The parametrization process of the negative binomial model is discussed later.

In general, for the inference of count data, the four most commonly used statistical model distributions are the Poisson, Negative Binomial, Hurdle, and Zero-Inflated regression models. The NB model addresses the issue of over-dispersion by including a dispersion parameter that relaxes the presumption of equal mean and variance

in the distribution whilst the Hurdle and Zero-Inflated regression models are utilized to handle the distribution of count outcome with excess zeroes[17–21].

The generalized linear model fails to consider the dependence of repeated observations over time. That means when data are measured repeatedly like CD4 counts of several individuals over time, the assumption of independence is no longer reasonable. Therefore, it is necessary to extend the GLM to generalized linear mixed-effects models, including a subject-specific random effect introduced in the *linear predictor* to seize the dependence.

Recall the linear mixed model:

$$y_{ij} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})X_{1ij} + \ldots + \left(\beta_p + b_{ip}\right)X_{pij} + \varepsilon_{ij},$$

where y_{ij} is an outcome variable, P is the predictor variable, $\beta_1, \ldots, \beta_p$ are fixed effects, $b_{i1}, \ldots, b_{ip}$ are random effects and ε_{ij}'s are residuals.

Suppose we want to generalize the above model. In that case, we do not need to assume that the outcome variable is normally distributed even after a transformation, such as the square root transformation for the CD4 count. However, it has to follow a distribution from the exponential family; at that point, we can combine the mixed model's idea with the generalized linear model. For instance, if y_{ij} is a count, we could look at Poisson regression. Hence the Poisson linear mixed model gets to be

$$\log\left(E\left(y_{ij}\right)\right) = \beta_0 + \beta_1 x_{1ij} + \cdots + \beta_p x_{pij} + b_0 + b_1 x_{1ij} + \cdots + b_p x_{pij}$$

In matrix notation form, the conditional mean of y_{ij} rely on fixed and random effects via the subsequent linear predictor:

$$\log\left\{E(y_{ij}|\boldsymbol{b_i})\right\} = \eta_{ij} = \boldsymbol{x}_{ij}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{'}\boldsymbol{b}_i.$$

where y_{ij}'s are independent and have a Poisson distribution, conditional on a vector of random effects b_i, with $var\left(y_{ij}|\boldsymbol{b_i}\right) = E\left(y_{ij}|\boldsymbol{b_i}\right)$, (*i.e.*, $\Phi = 1$), and $\boldsymbol{x}_{ij}^{'} = \boldsymbol{z}_{ij}^{'} = \left(1, t_{ij}\right)$. That is, the conditional mean of y_{ij} is associated with the linear predictor via a log link function, which is an example of a log-linear mixed-effects model[22,23].

Several methods are available to estimate the parameters (β_i's and b_i's) in GLMMs, which includes marginal quasi-likelihood (MQL), penalized (predictive) quasi-likelihood (PQL), the Laplace approximation, the Gauss-Hermite quadrative and the Markov Chain Monte Carlo (MCMC) method[24–27]. Our preference is for the Laplace approximation due to the fewer limitations than the Adaptive quadrature (method = quad). It is accurate, fast, and gives us the plausibility to use the likelihood and information criteria[26,28,29]. However, R-side random effects are not supported for method = laplace or method = quad in the Proc Glimmix statement. Instead, Proc Glimmix uses a random statement and the *residual option* to model repeated (R-side) effects.

"The parameter estimates based on the mixed-effects negative binomial model are not exceptionally different from those based on mixed-effects Poisson model. However, the Poisson model underestimates the SEs when over-dispersion is present, leading to improper inference. A straightforward way to select between these two models is to compare them based on a few criteria, such as AIC and BIC"[23]. Where for the ICs, a lower value means that the model fits better than the competing model. We may, moreover, compare models utilizing $-2loglikelihood$, and the *likelihood ratio test* for nested models. To some degree, parameters in GLMMs have different interpretations than parameters in the conventional marginal models. In GLMMs, the regression coefficients have subject-specific interpretations. Especially, they characterize the impact of variables on a particular subject's mean response. More specifically, the $\beta's$ are interpreted in terms of the effects of within-subject changes in explanatory variables on changes in an individual's transformed mean response, while holding the remaining covariates constant. Accordingly, β_j is interpreted as the change in an individual's log of response for a unit increase in $\boldsymbol{x_{ij}}$, while holding other fixed variables constant for that individual. Since the elements of the fixed effects, β_j, have interpretations conditional on b_i, the ith individual's random effects, they are regularly known as subject-specific regression coefficients. "Thus, GLMMs are most useful when the main scientific objective is to make inferences about individuals instead of the population average effects; the population averages are the targets of inference in marginal models"[22].

The negative binomial (NB) distribution, also the result of a Poisson–Gamma mixture, has vast applications as a model for count data, especially for data showing over-dispersion. It has properties that are comparable to the Poisson model, as discussed above, in which the outcome variable Y_i is modeled as a Poisson variable with a mean λ_i where the model error is assumed to follow a Gamma distribution. The Poisson-Gamma mixture model was developed to account for over-dispersion that is widely observed in discrete or count data[30]. The pdf of the NB distribution is frequently expressed in terms of the mean λ and dispersion parameter θ such that the probability of observing a non-negative integer k, which was given by Demidenko[31] parameterization of the negative binomial regression, discussed as follows:

If Y takes discrete values with the conditional Poisson distribution: $P_r(Y = k|\lambda) = \frac{e^{-\lambda}\lambda^k}{k!}$, where $\lambda > 0$, $\lambda \sim Gamma(\alpha, \theta)$ then the pdf of a two-parameter, α, *and* θ, Gamma distribution is given by:

$$f(\lambda; \alpha, \theta) = \frac{\lambda^{\alpha-1}e^{-\lambda/\theta}}{\theta^{\alpha}\Gamma(\alpha)}, \quad \lambda > 0, \quad \alpha > 0, \quad \theta > 0 \tag{2}$$

Thus, the negative binomial (Poisson–Gamma) model can be defined as:

$$f(Y|\lambda) = \frac{e^{-\lambda}\lambda^k}{k!}\frac{\lambda^{\alpha-1}e^{-\lambda/\theta}}{\theta^{\alpha}\Gamma(\alpha)} \tag{3}$$

It has also been defined in the literature as:

$$= \binom{\alpha+k-1}{k}\left(\frac{\theta}{1+\theta}\right)^{k}\left(\frac{1}{1+\theta}\right)^{\alpha} = \frac{\Gamma(\alpha+k)}{k!\Gamma(\alpha)}\left(\frac{\theta}{1+\theta}\right)^{k}\left(\frac{1}{1+\theta}\right)^{\alpha}, \tag{4}$$

where the binomial coefficient is computed as $\binom{\alpha+k-1}{k} = \frac{(\alpha+k-1)(\alpha+k-2)\ldots\alpha}{k!} = \frac{(\alpha+k-1)!}{k!(\alpha-1)!}$. Note that for a positive integer α, we have $\Gamma(\alpha) = (\alpha-1)!$.

For negative binomial distribution, $E(y) = \alpha\theta$, and $var(y) = \alpha\theta(1+\theta)$. For Poisson distribution, the mean and variance are equal, but the variance is higher than the mean by $\alpha\theta^2$ for negative binomial. By applying some calculus, one can show that the Poisson distribution is a special case of the negative binomial distribution when $\alpha \to \infty$ and $\theta \to 0$, such that the product, $\alpha\theta = \lambda$, is kept constant. The parameter a $= \frac{1}{\alpha}$ is associated with the "extra-Poisson" variation or over-dispersion because $var(y) = \lambda + a\lambda^2$, which is quadratic in the mean, that is why the negative binomial model is referred to as the NB2 model. This interpretation justifies a (λ, a) parameterization of the NB distribution as

$$P_r(Y = k;\ \lambda,\ \mathrm{a}) = \binom{k+\frac{1}{a}-1}{k}\left(\frac{a\lambda}{1+a\lambda}\right)^{k}\left(\frac{1}{1+a\lambda}\right)^{\frac{1}{a}},$$

where $E[y] = \lambda$ and $\mathrm{var}[y] = \lambda + a\lambda^2$, and $a = 0$ results in Poisson distribution. This latest parameterization is useful to specify the NB regression and for testing over-dispersion as $H_0 : a = 0$[32].

The likelihood function for Eq. (2) is proportional to

$$L(\boldsymbol{\beta}, \alpha) = \prod_{i=1}^{n} \frac{\Gamma(\alpha+k_i)}{k_i!\Gamma(\alpha)}\left(\frac{\theta_i}{1+\theta_i}\right)^{k_i}\left(\frac{1}{1+\theta_i}\right)^{\alpha}$$

Lawless[32] notes that for any $c > 0, \Gamma(k+c)/\Gamma(c) = c(c+1)\times\cdots\times(c+k-1)$ for integer-valued $k \geq 1$, thus, $\frac{\Gamma(\alpha+k)}{\Gamma(\alpha)} = \alpha(1+\alpha)\times\cdots\times(k-1+\alpha)$. Hence, $log\left\{\frac{\Gamma(\alpha+k)}{\Gamma(\alpha)}\right\} = \sum_{j=0}^{k_i-1}\log(\alpha+j)$. This produces $\log L(\boldsymbol{\beta}, \alpha)$ as follows

$$= \sum_{i=1}^{n}\left(\sum_{j=0}^{k_i-1}\log(\alpha+j) - \log k_i! + k_i\log\theta_i - k_i\log(1+\theta_i) + \alpha\log 1 - \alpha\log(1+\theta_i)\right)$$

$$\ell(\boldsymbol{\beta}, \alpha) = \sum_{i=1}^{n}\left(\sum_{j=0}^{k_i-1}\log(\alpha+j) - \log k_i! + k_i\log\theta_i - (k_i+\alpha)\log(1+\theta_i)\right)$$

Therefore, applying the Poisson theorem with Gamma distribution leads to the negative binomial distribution. Furthermore, detailed discussions of estimating methods and characteristics of the negative binomial model are presented in numerous literature[13,14,25,30–32].

When repeated counts are measured on the same individual over time, the assumption of independence is no longer reasonable; instead, they are correlated. Subject-specific random effects can be added into the linear predictor to modeling such dependence. Let y_{ij} be the values of a count variable (non-negative integer value) for subject i at time point j. The count is assumed to be drawn from a Poisson distribution with errors assumed to have a normal distribution, $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$. Then, the Poisson mixed-effects model that specifies the expected number of counts is written as

$$\log(\mu_{ij}) = \boldsymbol{x}_{ij}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{'}\boldsymbol{b}_i + \varepsilon_{ij}, \tag{5}$$

where $\boldsymbol{x}_{ij}$ is the variable of interest, $\boldsymbol{\beta}$ is the vector of fixed effects (population-level effects), including an intercept β_0, $\boldsymbol{b}_i$ is the vector of random effects (subject-level effects) for the sample variables $\boldsymbol{z}_{ij}$, and ε_{ij} is the random errors[22,23]. Given the Poisson process for the count y_{ij}, the probability that $y_{ij} = y$, conditionally on the random effects $\boldsymbol{b}_i$, is given by

$$P(y_{ij} = y|\boldsymbol{b}_i, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}) = \frac{e^{-\mu_{ij}}\mu_{ij}^{y}}{y!} = \frac{1}{y!}e^{-\exp\left(\boldsymbol{x}_{ij}^{'}\boldsymbol{\beta}+\boldsymbol{z}_{ij}^{'}\boldsymbol{b}_i\right)}\exp\left(\boldsymbol{x}_{ij}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{'}\boldsymbol{b}_i\right)^{y}$$

$$= \frac{1}{y!}\exp\left[\left(\boldsymbol{x}_{ij}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{'}\boldsymbol{b}_i\right)^{y} - \exp\left(\boldsymbol{x}_{ij}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{'}\boldsymbol{b}_i\right)\right], y = 0, 1, 2, \ldots$$

This addition also can be applied to the NBMM that allows over-dispersion by assuming a gamma distribution for the errors; instead of a normal distribution. Suppose that $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ are known vectors of covariates associated with count data y_{ij}, $\quad i = 1, \ldots, n \quad$ and $\quad j = 1, \ldots, n_i$, conditional on a $q-$ dimensional vector of subject-specific random effects, $\boldsymbol{b_i}$, the counts of y_{ij}, with the assumption of gamma errors, has a negative binomial distribution, $y_{ij}|\boldsymbol{b_i} \sim NB\left(\mu_{ij}, \mu_{ij} + \theta\mu_{ij}^2\right)$, with $\mu_{ij} = E(y_{ij}|\boldsymbol{b_i}) = \exp\left\{\boldsymbol{x}_{ij}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{'}\boldsymbol{b}_i\right\}$. This indicates that the mean parameters μ_{ij} of the negative binomial mixed-effects models are also related to the predictor variables $\boldsymbol{x}_{ij}$, and the sample variables $\boldsymbol{z}_{ij}$ through the logarithm link function: $\log(\mu_{ij}) = \boldsymbol{x}_{ij}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{'}\boldsymbol{b}_i + \varepsilon_{ij}$, which shows

Covariates	Level	CD4 count N (%)			p-value	% Missing
		< 200	200–500	> 500		
Baseline BMI category	Underweight	2 (0.03)	219 (3.12)	254 (3.62)	< 0.0001	0.0
	Normal weight	114 (1.62)	2305 (32.84)	2690 (38.32)		
	Overweight	18 (0.26)	512 (7.29)	657 (9.36)		
	Obese	0	17 (0.24)	231 (3.29)		
Baseline viral load	Undetected	0	0	16 (0.23)	< 0.0001	0.0
	Low	20 (0.28)	791 (11.27)	1532 (21.83)		
	Medium	45 (0.64)	1209 (17.22)	1497 (21.23)		
	High	69 (0.98)	1053 (15)	787 (11.21)		
Number of sexual partners	No partner	29 (0.41)	565 (8.05)	579 (8.25)	< 0.0001	0.0
	Stable partner	85 (1.21)	2274 (32.4)	3078 (43.85)		
	Many partners	20 (0.28)	214 (3.05)	175 (2.49)		
Age group	< 20	1 (0.01)	130 (1.82)	121 (1.72)	< 0.0001	0.0
	20–29	97 (1.38)	1872 (26.67)	1977 (28.17)		
	30–39	17 (0.24)	813 (11.58)	1255 (17.88)		
	40–49	19 (0.27)	203 (2.89)	369 (5.26)		
	50–59	0	35 (0.5)	91 (1.3)		
	≥ 60	0	0	19 (0.27)		
Educational level	Primary school	3 (0.04)	104 (1.48)	181 (2.58)	0.0129	0.0
	Secondary school	131 (1.87)	2949 (42.01)	3651 (52.02)		
Place of residence	Rural	62 (0.88)	1467 (20.90)	1806 (25.73)	0.7176	0.06
	Urban	72 (1.03)	1586 (22.6)	2026 (28.86)		
ART initiation group	Pre ART	110 (1.57)	2566 (36.56)	2783 (39.65)	< 0.0001	0.0
	Post ART	20 (24)	487 (6.94)	1049 (14.95)		

Table 1. Distribution of CD4 count and associated selected covariates with percent missing. The response variable (CD cell count) has 110 (1.5%) missing observations.

that the model for the conditional mean of the NBMM is similar to that of PMM. However, the conditional variance of y_{ij} for NBMM is $Var(y_{ij}|\boldsymbol{b_i}) = \mu_{ij} + \theta\mu_{ij}^2$, which is greater than the conditional mean of PMM by $\theta\mu_{ij}^2$, specifically, because a gamma distribution is assumed for the exponentiated errors, $\exp(\varepsilon_{ij})$, with a mean of 1 and variance θ[22,31]. Random effects are used to demonstrate multiple assets of variations and subject-specific effects. As a result, they avoid biased inference on the fixed effects. The random effects are assumed to have a multivariate normal distribution:

$$\boldsymbol{b}_i \sim N(0, \Psi) \tag{6}$$

where Ψ is a positive-definite variance–covariance matrix that accounts for the correlation of the random effects[33,34].

Ethics approval and consent to participate. Ethical approval for the study was obtained from the Research Ethics Committee of the University of KwaZulu-Natal (E013/04), the University of the Witwatersrand (MM040202), and the University of Cape Town (025/2004). All participants provided written informed consent. All methods were performed following the relevant guidelines and regulations expressed in the Declaration of Helsinki.

Results

Table 1 shows the summary of CD4 count and its associated selected covariates in the CAPRISA 002 AI Study. The dataset included 235 subjects (7129 observations consists of a minimum of two and a maximum of sixty-one observations per subject). P-values demonstrated in Table 1 are obtained from the Chi-square test. At a 5% level of significance, the univariate cross-tabulation analysis uncovers that the patient's baseline BMI, baseline VL, number of sexual partners, age, ART initiation, and education level are significantly associated with patient's CD4 count. Table 1 demonstrates that there is a high prevalence of CD4 count above 500 cells/mm^3 among patients with normal weight and overweight status, which are 38.32 and 9.36%, respectively (p-value < 0.0001). Out of 7129 observations, patients with an undetectable viral load at baseline indicate no sign of a CD4 count < 500 cells/mm^3 throughout the study.

Moreover, from Table 1, there is a high prevalence of CD4 count above 500 cells/mm^3 for patients with low viral load at baseline (21.83%). This shows ART suppresses the amount of HIV viably in patient's body fluids who have an undetectable and low viral load at baseline to the point where standard tests are incapable of detecting any HIV or can only find a little flow. There is also a high prevalence of CD4 count above 500 cells/mm^3 for patients with a stable sexual partner (43.85%, p-value < 0.0001) compared to patients who have many sexual partners. A high prevalence of CD4 count above 500 cells/mm^3 is observed among patients of the age group between

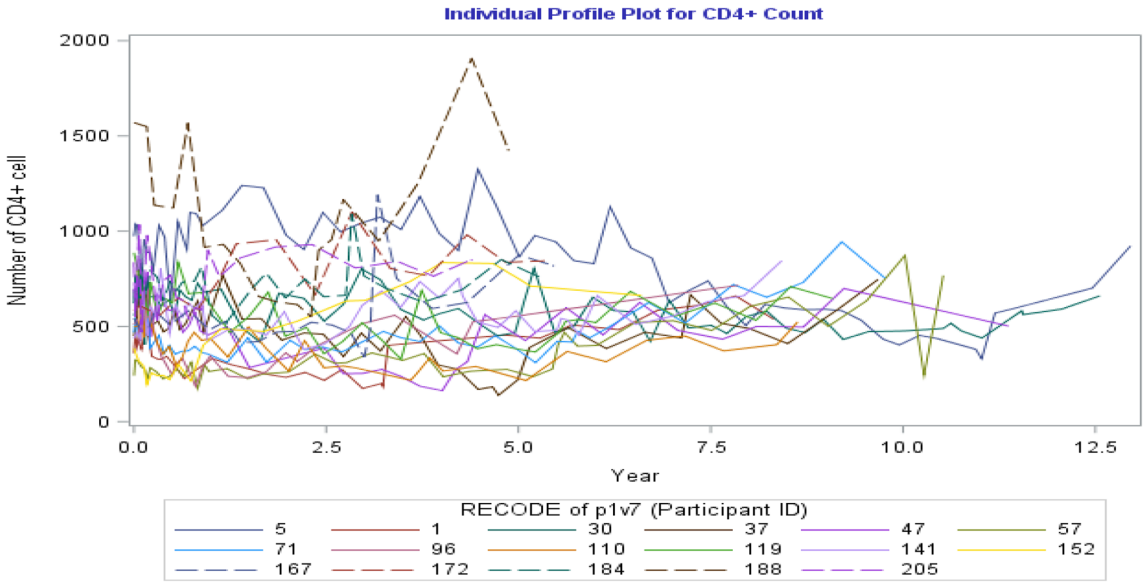


Figure 1. Individual profiles plot of CD4 cell count for 17 randomly selected subjects.

Distribution	Fit statistics					
	− 2 log likelihood	AIC	AICC	BIC	CAIC	HQIC
Poisson	204,842.9	204,892.9	204,893.1	204,979.4	205,004.4	204,927.8
NB	87,781.28	87,833.28	87,833.48	87,923.23	87,949.23	87,869.54

Table 2. Comparisons of fit statistics for the two distributions.

Fit Statistics for Conditional Distribution	Poisson	NB
− 2 log L(CD4 counts/r. effects)	199,670.3	85,320.39
Pearson χ^2	145,017.0	6396.89
Pearson χ^2/DF	20.66	**0.91**

Table 3. Measure of over-dispersion between Poisson and negative binomial distribution.

20–29 years and 30–39 years, which are 28.17 and 17.88%, respectively (p-value < 0.0001). The prevalence of CD4 count above 500 cells/mm^3 is also observed among women patients who have higher/secondary school levels of education (52.02%, p-value = 0.0129). However, the place of residence is found not to be associated with patients' CD4 count (p-value = 0.7176).

The individual profiles plot for 17 randomly selected HIV-Infected women enrolled in the CAPRISA 002 AI Study is shown in Fig. 1.

Analyzing data shown in Fig. 1, we can observe insights concerning the variability among individual patients at a given point in time, the variance within units over time, and the trends over time. Note that the space between the lines represents between unit variability, and the change in each line (slope) represents within variability. Moreover, as portrayed in Fig. 1, the number of CD4 cells seems to represent a slightly increasing pattern over time; however, the rate of increment is low. Additionally, Fig. 1 shows that there is wide variability in the number of CD4 cells and in the number of repeated measures (number of observations per subjects are not equal).

The results of the Fit statistics in Table 2 are obtainable because of method = Laplace in Proc Glimmix Procedure. These values are relative and valuable when we compare different model choices. The NB model's Fit statistics are much smaller than the Poisson model (Table 2). For instance, AICC is 87833.48 for NB versus 204893.1 for the Poisson. Also, the Pearson χ^2/DF of 20.66 for the Poisson model is problematic (Table 3), indicating evidence of over-dispersion in the data. Ideally, this value ought to be generally 1.0 when modeling count data with a Poisson distribution. The ratio of Pearson Chi-Square statistics is dropped from 20.66 to 0.91 under the NB model, which is close to one (Table 3), indicating that over-dispersion has been appropriately modeled and it is no longer an issue under the NB model.

In addition to the conditional fit statistics, any other diagnostic that may allow us to see over-dispersion in the Poisson model is a graphical representation (Fig. 2). We can get residual plots through Proc Glimmix using the Plot option. Here, we only focus on looking at residual versus predicted plots. Figure 2 (left panel) shows

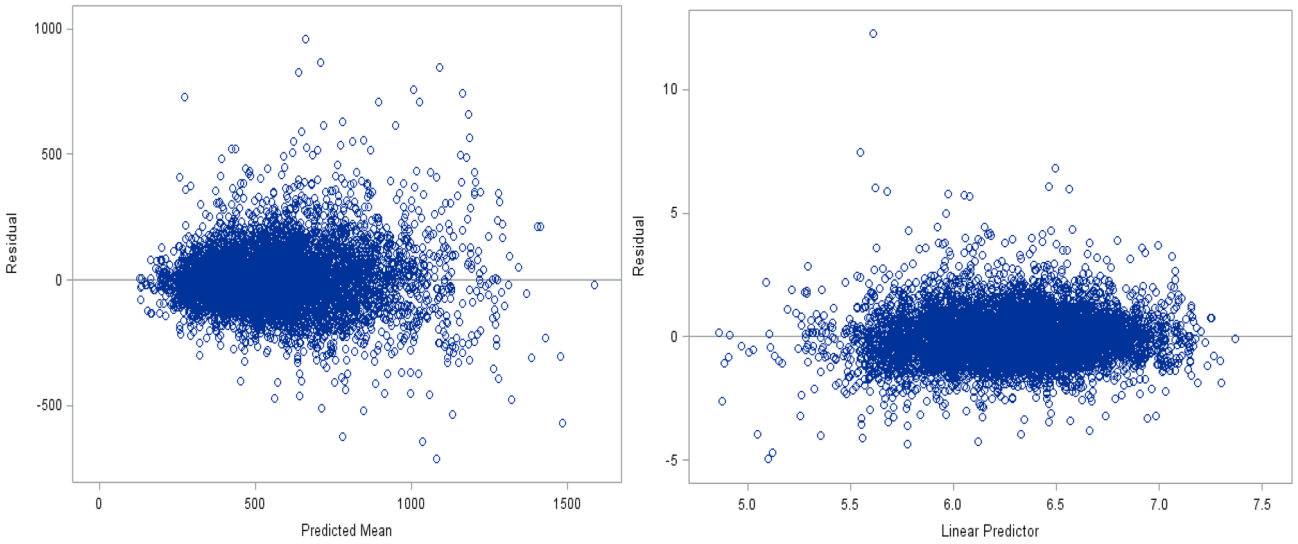


Figure 2. Data-scale raw residuals and Model-scale studentized residuals versus predicted values.

Random effect models	Information criteria					
	− 2log ℓ	AIC	AICC	BIC	CAIC	HQIC
Model 1	87,781.28	87,833.28	87,833.48	87,923.23	87,949.23	87,869.54
Model 2	88,603.50	88,649.50	88,649.66	88,729.07	88,752.07	88,681.58
Model 3	88,591.64	88,637.64	88,637.80	88,717.21	88,740.21	88,669.72
Model 4	89,156.39	89,202.39	89,202.55	89,281.96	89,304.96	89,234.47
Model 5	89,837.18	89,879.18	89,879.31	89,951.83	89,972.83	89,908.47
Model 6	92,302.08	92,344.08	92,344.21	92,416.73	92,437.73	92,373.37
Model 7	91,190.61	91,232.61	91,232.74	91,305.26	91,326.26	91,261.90

Table 4. Comparison of random effect models.

the visual prove of over-dispersion. As the Predicted Mean ($\hat{\mu}$) increases, the associated residuals become more broadly dispersed. The variance ought to increase as a function of the mean, but not as quickly as we see in this plot (Fig. 2). Also, Fig. 2 (right panel) shows prove of over-dispersion. The variance adjusted residuals are more variable around the lower point of the estimated Linear Predictor ($\hat{\eta}$). On the model scale (Fig. 2 (right panel)), we should not see the variance adjusted residuals variable across different points of $\hat{\eta}$ as we see in this plot[16,35]. In other words, Fig. 2 (right panel) demonstrates that the empirical distribution of the residuals is not reasonably symmetric, and in general, it is not very informative.

The improvement in the Pearson χ^2/DF and Fit statistics indicate that it is best to model data from this experiment with the NB distribution. Utilizing the proper distribution gives unbiased test statistics and SE estimates (Table 4).

In addition, the subsequent random effect models were taken into consideration for testing NBMMs:

Model 1: *Intercept, Time,*$\sqrt{Time}$.
Model 2: *Intercept, Time.*
Model 3: *Intercept,* $\sqrt{Time}$.
Model 4: *Time,* $\sqrt{Time}$.
Model 5: *Intercept only.*
Model 6: *Time only.*
Model 7: $\sqrt{Time}$ *only.*

We conclude that Model 1 is a preferable model among models listed above since it has the smallest information criteria. Moreover, a comparison of the covariance structure using the fitted model (Supplementary Table S1) and a comparison of fixed-effects results across different covariance structures using Model 1 (Supplementary Table S2) are made. The estimated unstructured covariance matrix ($\hat{D}$) for the GLMMs model that uses NB distribution is

Effect	Num DF	Den DF	NB		Poisson	
			F value	Pr > F	F value	Pr > F
Time in month	1	235	62.53	< 0.0001	14.80	0.0002
Sqrt_Time	1	234	86.36	< 0.0001	48.41	< 0.0001
Baseline BMI category	3	6307	6.26	0.0003	6.31	0.0003
ART initiation	1	6307	345.45	< 0.0001	5890.28	< 0.0001
Baseline VL	3	6307	7.48	< 0.0001	12.79	< 0.0001
No. of sexual partners	2	6307	1.64	0.1935	1.85	0.1578
Age group	5	6307	1.46	0.1987	27.34	< 0.0001
Education level	1	6307	0.25	0.6196	0.15	0.6990
Place of residence	1	6307	0.01	0.9246	0.11	0.7406

Table 5. Type III Analysis of fixed effects for Poisson and NB distribution.

Covariates	Negative binomial mixed-effects model				Poisson mixed-effects model		
	Estimate	SE	Pr > |t|	95% CI for NB estimate	Estimate	SE	Pr > |t|
Intercept	6.4697	0.04982	< 0.0001	(6.3715, 6.5679)	6.4625	0.04264	< 0.0001
Time in month	0.007824	0.000989	< 0.0001	(0.005875, 0.009774)	0.006564	0.001706	0.0002
Sqrt_Time	− 0.08649	0.009307	< 0.0001	(− 0.1048, − 0.06815)	− 0.06839	0.009830	< 0.0001
ART initiation (post)	0.2301	0.01238	< 0.0001	(0.2058, 0.2543)	0.1947	0.002537	< 0.0001
Baseline BMI category (ref. = normal weight)							
Obese	0.4815	0.1113	< 0.0001	(0.2633, 0.6996)	0.4985	0.1147	< 0.0001
Overweight	0.02561	0.04975	0.6067	(− 0.07191, 0.1231)	0.03131	0.05148	0.5431
Underweight	0.005901	0.07927	0.9407	(− 0.1495, 0.1613)	0.01691	0.08264	0.8379
Baseline HIV viral load category (ref. = low VL)							
High VL	− 0.2393	0.05157	< 0.0001	(− 0.3404, − 0.1382)	− 0.3074	0.05065	< 0.0001
Medium VL	− 0.1258	0.04587	0.0061	(− 0.2157, − 0.03585)	− 0.1121	0.04686	0.0168
Undetectable	0.1377	0.2901	0.6351	(− 0.4310, 0.7064)	0.1199	0.2978	0.6872
Number of sexual partners (ref. = stable partner)							
Many partners	− 0.1560	0.09394	0.0967	(− 0.3402, 0.02811)	− 0.1674	0.09908	0.0911
No partner	− 0.04821	0.04993	0.3343	(− 0.1461, 0.04967)	− 0.05913	0.05164	0.2522
Age group in years (ref. = < 20)							
20–29	0.01166	0.03104	0.7072	(− 0.04919, 0.07251)	− 0.00791	0.007830	0.3125
30–39	0.02852	0.03432	0.4060	(− 0.03876, 0.09580)	− 0.01239	0.008474	0.1438
40–49	− 0.00719	0.04545	0.8743	(− 0.09629, 0.08191)	− 0.03422	0.01112	0.0021
50–59	− 0.05694	0.06662	0.3927	(− 0.1875, 0.07365)	− 0.1399	0.01549	< 0.0001
≥ 60	0.2082	0.1532	0.1741	(− 0.09205, 0.5084)	− 0.3107	0.03519	< 0.0001
Education attainment (ref. = secondary or high school)							
Primary school	− 0.04509	0.09084	0.6196	(− 0.2232, 0.1330)	− 0.03582	0.09263	0.6990
Residence of participant (ref. = urban)							
Rural	− 0.00373	0.03947	0.9246	(− 0.08112, 0.07365)	0.01337	0.04038	0.7406

Table 6. Parameter estimates using Poisson and NB mixed-effects model.

$$\hat{D} = \begin{bmatrix} 0.1131 & 0.000739 & -0.01754 \\ 0.000739 & 0.000155 & -0.00137 \\ -0.01754 & -0.00137 & 0.01556 \end{bmatrix}$$

The estimated scale parameter is 0.04205, which can be found in the "Covariance Parameter Estimates" output of the SAS PROC GLIMMIX (Laplace) procedure (see Supplementary Table S3). Therefore, the estimated conditional variance of the count is $\hat{\mu}_i + 0.04205\hat{\mu}_i^2$, where $\hat{\mu}_i$ is the conditional mean on the counting scale. "The Scale parameter measures the magnitude of over-dispersion and is practically equivalent to the mean square error in conventional theory analysis of variance"[15].

Table 5 shows the overall effect of the selected factors within the fitted models. The results indicate that the effects of Time, Baseline BMI, HAART initiation group, baseline viral load, and the number of sexual partners on the patient's CD4 count were found to be highly significant in both fitted models. However, the overall F-values

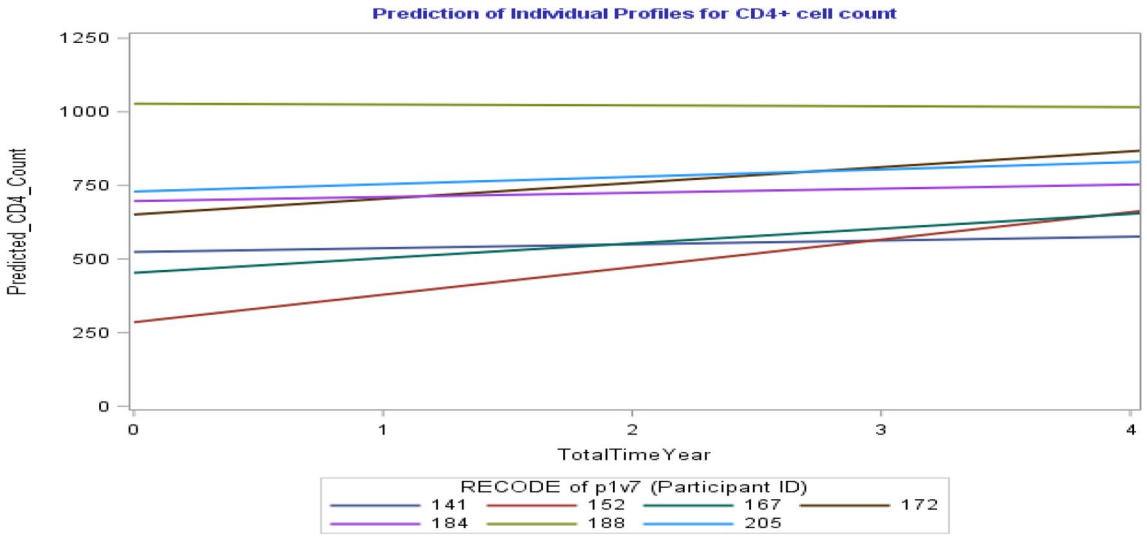


Figure 3. Prediction of 7 randomly selected individual profiles plot of CD4 count for 4 years.

of the NB model were smaller than for the Poisson model. This can be supporting prove that over-dispersion can lead to inflated and biased F-values if we do not use the proper model in our analysis.

Table 6 shows the log of the expected CD4 count as a function of the selected predictor variables using a negative binomial mixed-effect model. The results indicate that time (month) significantly affects the CD4 count of a patient. We interpret the coefficient of the month as an average within-subject change in the logs of expected CD4 count for patients would be expected to increase by 0.0078 units (p-value < 0.0001; 95% CI 0.005875, 0.009774), while holding other factors in the model constant. The square root of time shows a significant adverse effect in the logs of expected CD4 counts of a patient (Table 6). Compared to pre HAART initiation, the difference in the logs of CD4 counts of a patient who had been initiated on HAART would be expected to increase by 0.2301 units (p-value < 0.0001; 95% CI 0.2058, 0.2543), holding other factors constant in the model. It can be observed that the difference in the logs of expected CD4 counts is expected to be 0.4815 units (p-value < 0.0001; 95% CI 0.2633, 0.6996) higher for patients with higher BMI (Obese) at baseline compared to patients with normal weight status holding other factors constant in the model. Those patients who had high and medium viral load at baseline, the difference in the logs of their expected CD4 counts were decreased by 0.2393 (p-value < 0.0001; 95% CI − 0.3404, − 0.1382) and 0.1258 (p-value = 0.0061; 95% CI − 0.2157, − 0.03585), respectively, compared to patients who had low viral load at baseline while holding other factors in the model constant.

Furthermore, the SEs for the Poisson mixed-effects model were more likely to be underestimated and/or biased compared to those from a negative binomial mixed-effects model since the model is fitted by ignoring over-dispersion of the data (Table 6).

The prediction profile equation for the average number of CD4 cell following Table 6 results obtained by NB mixed-effects model is given as:

$$\begin{aligned} log\left(\hat{\mu}_i\right) = {} & 6.4697 + 0.007824 \times time - 0.08649 \times \sqrt{time} + 0.2301 \\ & \times postHAARTtreatment + 0.4815 \times obese - 0.2393 \\ & \times highVL - 0.1258 \times mediumVL. \end{aligned}$$

Taking antilog values on both sides of the above-predicted equation yields the expected number of counts, given by

$$\begin{aligned} \hat{\mu}_i = \exp\Big(& 6.4697 + 0.007824 \times time - 0.08649 \times \sqrt{time} + 0.2301 \times postHAARTtreatment \\ & + 0.4815 \times obese - 0.2393 \times highVL - 0.1258 \times mediumVL \Big). \end{aligned}$$

The prediction of individual profiles, Fig. 3, presents the estimated trajectories for the average number of CD4 cell under the estimates acquired by the negative binomial mixed-effect model with UN covariance structure consolidated with the model where the intercept and slope were considered as random effects (see Table 4 and Supplementary Table S1) for seven patients with particular profiles for four years. For instance, from CAPRISA 002 AI Study, patient ID = 141, 22 years old female, with around 500 cells/mm^3 CD4 cell count at baseline, low VL at baseline, had normal weight status at baseline, and have no sexual partner at the time of enrollment.

The second patient ID = 152, 34 years old female, with obese weight status at baseline, having stable sexual partner, high VL at baseline, and CD4 count at baseline below 500 cells/mm^3. As a third example, we looked at patient ID = 172 who had undetected VL at baseline, with CD4 count at baseline above 500 cells/mm^3, 29 years old female, with obese weight status at baseline and have a stable sexual partner. As a fourth example, we can also look at patient ID = 188, who had a high number of CD4 cells at baseline (1070 cells/mm^3) with low VL at baseline, 42 years old, had obese weight status at baseline, and have a stable sexual partner. As we would anticipate,

Parameter	Parameter estimates (10 imputations)					
	Estimate	SE	Pr>|t|	95% confidence limits	Minimum	Maximum
Intercept	6.459413	0.049830	< 0.0001	(6.36175, 6.55708)	6.458658	6.460775
Time in month	0.007475	0.000975	< 0.0001	(0.00556, 0.00939)	0.007450	0.007508
Sqrt_Time	− 0.083647	0.009266	< 0.0001	(− 0.10181, − 0.06549)	− 0.083982	− 0.083434
ART initiation (Post)	0.224037	0.012594	< 0.0001	(0.19935, 0.24872)	0.223216	0.225014
Baseline BMI category (ref. = normal weight)						
Obese	0.474714	0.109902	< 0.0001	(0.25931, 0.69012)	0.473892	0.475630
Overweight	0.024208	0.048971	0.6211	(− 0.07177, 0.12019)	0.023820	0.024529
Underweight	0.002070	0.078101	0.9789	(− 0.15101, 0.15515)	0.001321	0.003137
Baseline HIV viral load category (ref. = Low VL)						
High VL	− 0.239102	0.051294	< 0.0001	(− 0.33964, − 0.13857)	− 0.239735	− 0.238839
Medium VL	− 0.122078	0.045390	0.0072	(− 0.21104, − 0.03311)	− 0.122251	− 0.121642
Undetectable	0.142848	0.286259	0.6178	(− 0.41821, 0.70391)	0.142510	0.143351
Number of sexual partners (ref. = stable partner)						
Many partners	− 0.153632	0.092090	0.0953	(− 0.33412, 0.02686)	− 0.154667	− 0.152911
No partner	− 0.046962	0.049227	0.3401	(− 0.14344, 0.04952)	− 0.047267	− 0.046691
Age group in years (ref. = < 20)						
20–29	0.013477	0.031659	0.6703	(− 0.04857, 0.07553)	0.012306	0.014325
30–39	0.033725	0.034974	0.3349	(− 0.03482, 0.10227)	0.032678	0.034744
40–49	− 0.005842	0.046177	0.8993	(− 0.09635, 0.08466)	− 0.007790	− 0.004745
50–59	− 0.052070	0.067501	0.4405	(− 0.18437, 0.08023)	− 0.054207	− 0.051024
≥ 60	0.206708	0.156046	0.1853	(− 0.09914, 0.51255)	0.205360	0.207553
Education attainment (ref. = secondary or high school)						
Primary school	− 0.046292	0.089605	0.6054	(− 0.22191, 0.12933)	− 0.046602	− 0.046009
Residence of participant (ref. = urban)						
Rural	− 0.001916	0.038813	0.9606	(− 0.07799, 0.07416)	− 0.002146	− 0.001596

Table 7. Combined results of a negative binomial mixed-effects model analysis using MI Procedure to deal with the missing values.

all seven individuals appeared to have an increased average number of CD4 cells over time, in line with their predicted individual profiles (Fig. 3). However, the increasing level or degree is different among individuals. This is due to factors related to this study and numerous other characteristics of these individuals, mainly (according to our research) for their VL at baseline, baseline BMI and the treatment (either the patient had effective HAART initiation after HIV exposure or not).

Moreover, for this study to yield meaningful results, we checked the missing values in the dataset using the Little's MCAR test. The regular Little's MCAR test gives us a χ^2 distance of 4515.686 with a degree of freedom 106 and p-value 0.000 (Little's MCAR test: Chi-Square = 4515.686, DF = 106, sig. = 0.000). The analysis gives evidence that the missing data in the study variables of interest are not MCAR under significance level 0.000. Therefore, we used Multiple Imputation (MI) techniques to get a valid analysis for parameter estimates from the complete data set by fitting the chosen model. The MI procedure's main concept is to replace each missing value with a set of *m* possible values. Generally, the imputation of dependent and independent variables is basic for getting unbiased estimates of the regression coefficients[36]. Following Rubin's (1987) terminology, the MI procedure includes three distinct phases: each missing value is imputed *m* times to generate *m* complete data sets, analyze each *m* complete data sets separately by using standard procedure and then combine the results to generate valid statistical inference about the model parameters from the *m* data set analysis using Rubin's combine rule[37]. SAS Proc MI can be used to create *N* number of imputation; after that, Proc MIAnalyze is used to pool the parameter estimates. A detailed discussion of missing data analysis and how missing data handled by statistical software can be found in numerous literature[37–44].

Table 7 shows a combined result for each parameter. The table also shows a 95% confidence interval, the minimum and maximum regression coefficients from the imputed data set, and the associated p-value. We can compare the results given in Table 7 with the results of applying the negative binomial mixed-effect model to the CAPRISA 002 AI data using incomplete cases (Table 6). Comparing the two different sets of results, we do not see that many exciting differences. In both cases, covariates that were found to be significantly affecting the patient's CD4 count are similar, and their respective parameter estimates are more close to each other.

In general terms, a comparison of the results from data with missing value case analysis (Table 6) and multiple imputation analysis (Table 7) shows little difference between parameter estimates, SEs, and confidence intervals. In this case, the small difference in results and associated inferences is likely due to relatively low amounts of missing data in the analysis variables (Table 1). However, it will not always be true that results from incomplete or complete case analysis and a multiple imputation treatment of the data will lead to similar results and inferences[38]. Finally, missing data is especially common in longitudinal data sets. Missingness can arise due to respondent

attrition, survey structure, file-matching issues, and refusal to answer sensitive questions such as certain health conditions, illegal behaviors, or income[38]. Missing data can also arise due to death. A loss to follow-up due to death is qualitatively different from dropout due to other responses and, ordinarily, needs to be dealt with quite differently in the analysis of longitudinal data[9]. Missing data is generally classified as Missing Completely at Random (MCAR), Missing at Random (MAR), or Not Missing at Random (NMAR)[37,39,41,44–46].

Discussion and conclusion

GLMs extend the standard concept of linear models to outcome variables whose distribution is from a member of the exponential family. "GLM consists of three components: a *stochastic* component that characterizes the likelihood distribution of the response variable; a *linear predictor* that is a *systematic* component portraying the linear model characterized by the explanatory variables; and a *link function* that connect the mean of the response variable to a linear combination of the explanatory variables. Link functions that are commonly used for distributions are discussed in numerous literature"[12,16,24,28,35,47–51]. Parameters in GLM are estimated based on maximum likelihood principles. Different ways of transformations of the response variable make the transformed data to fulfill the linear model's assumptions, such as approximately normally distributed and having stable variances. In a more common term, a transformation is a replacement that changes the shape of distribution or relationship. However, transformation is often challenging for regression settings in which it additionally influences the practical relationship between the covariates and the outcome variable. In some cases, it is not perceived that the utilization of transformations changes the model[52].

Transformations are elaborative when a selected choice is not predetermined through different considerations; that is, the selection of transformation is subjective[53]. "GLMs avoid these problems since the data are no longer transformed; instead, a function of the means is modeled as a linear combination of the covariates"[24,48]. Sometimes, for example, for large values of the estimated coefficient, the use of a transformation is effective than using GLMs and Wald type statistics for inference[48,49]. "In general, however, transformations rarely compete well with GLMs for adequately powered studies"[48]. Therefore, we analyzed the non-normal untransformed form of the CD4 cell count of a patient enrolled in the CAPRISA 002 AI Study in the context of GLMMs (Table 6).

Longitudinal studies, also called mixed-effects models, are used to study changes in the response variable over a relevant interval of time or space and the effects of different factors on these changes. The two fundamental issues in longitudinal studies are constructing an appropriate model for the mean and choosing a reasonable but parsimonious model for the covariance structure of longitudinal data[22]. For these reasons, we have fitted an NBMM consolidated with the UN covariance structure since there was enough evidence of over-dispersion in the data. The chosen covariance structure gives the smallest information criteria (Supplementary Table S1). The comparisons between Poisson and negative binomial mixed-effects models were outlined in Table 6.

Moreover, comparisons of the covariance structure illustrated in Supplementary Table S1. GLMMs combine the GLMs with the LMMs. "As an extension of GLMs, they consolidate random effects into the linear predictor. As a mixed model, they contain at least one fixed effect and at least one random effect"[54]. Parameter estimation in GLMMs is also based on maximum likelihood principles; inferences for the parameters are readily obtained from classical maximum likelihood theory[22,54]. "The two fundamental computational methods to attain solutions to the likelihood equations are a *pseudo-likelihood*, and integral approximation of the log-likelihood using either the Laplace or Gauss-Hermite quadrature strategies"[16,40,55]. Since *pseudo-likelihood* generates biased covariance parameter estimates when the number of observations per subject is small, it is especially inclined to biased estimates when the power is small and uses a *pseudo-likelihood* rather than a true likelihood, likelihood ratio, and fit statistics such as AICC and BIC have no clear meaning. However, the integral approximation uses the actual likelihood and grant us the appropriate likelihood ratio tests or information criteria, permitting competing models to be compared using these test statistics. Of these two, the Laplace method is best since quadrature is ordinarily computationally restrictive for regularly repeated measures. Moreover, the Laplace procedure is less computationally intensive than the quadrature procedure and is considerably more flexible in terms of the models with which it can be used. Detailed discussions of parameter estimation in GLMMs can be found in numerous literature[16,22,28,47,48,51]. The fit statistics in Table 3 were obtained by using the Laplace method. If this method had not been specified on the SAS Proc Glimmix procedure, the default *pseudo-likelihood* method would have been used to fit the model. Because *pseudo-likelihood* is based on Tylor series approximation to the conditional likelihood and not expressly on the conditional likelihood itself, a goodness of fit statistic which includes the Pearson χ^2 that is particularly appropriate to the conditional distribution cannot be computed. Rather, the *pseudo-likelihood* approaches calculate a Generalized χ^2 statistic that measures the combined fit of the conditional distribution of the counts and the random effects. Since it is not particular to solely the conditional distribution, it does not offer a clear cut diagnostic to evaluate the fit of the Poisson distribution to the counts[40].

The Pearson χ^2/DF gives the goodness of fit statistic to evaluate over-dispersion within the Poisson model. Since the variance and mean of the Poisson are equal, the scale parameter (α) is 1. If the Poisson assumption is fulfilled, the Pearson χ^2/DF ought to be close to 1. Its estimated value of 20.66 (Table 3) indicated solid prove of over-dispersion under the Poisson model. "Over-dispersion would mean more variability shown by the data than would be assumed under a given statistical model"[20]. Over-dispersion could be an issue that should not be disregarded in the statistical inferences. The essential and most critical outcome of over-dispersion is its effect on SEs and test statistics. This was demonstrated in Table 5, uncorrected analysis of over-dispersed data (Poisson model) consequences underestimated SEs, leading to biased estimates and inflated test statistics. "It is basic to check for over-dispersion when fitting a GLM or a GLMM to guarantee that inferences derived from the fitted model are precise"[20]. Over-dispersion is an implication that the fitted model is incorrect, and adjustments are required. "The two most commonly used approaches in GLMMs, to avoid unwanted outcomes outlined above, are: adjusting the SEs and test statistics by incorporating an adjustment for over-dispersion in the model or

assume a different probability distribution for the counts that more reasonably approximate the method by which over-dispersion emerge"[48]. Because the second strategy of assuming a different distribution is a reasonable and suggested methodology, it was illustrated in Table 5 in which the negative binomial distribution substitutes the Poisson distribution as the conditional distribution of the outcome. The NB distribution is the foremost candidate as an alternative to the Poisson[13,14]. The Pearson χ^2/DF value of 0.91 (Table 3) shows that the negative binomial gives a much-improved fit of the data compared to the Poisson model. This is one of a reasonable GLMMs approach for managing with over-dispersion.

Supplementary Table S2 outlined that the fixed effects are significantly influenced by the covariance structure. Furthermore, the covariance structure also impacted the random effects estimate: the time effects and their SEs. The SEs tend to be affected more than the estimates. The selection of covariance structures subjects for non-normally distributed data, just as it does for normally distributed data. The fit statistics related to *pseudo-likelihood* estimation are not comparable among models. Consequently, the fit statistics cannot be used to select between competing for covariance structures. Therefore, the choice of covariance structure is not as straightforward for non-normal longitudinal response data as it is under normality assumption[15,52,55–58]. However, for the GLMM approach, the situation is better. As we discussed previously, since the GLMM characterizes an exact probability process under the Laplace method, fit statistics such as AICC and BIC can be obtained[57]. Thus, for GLMMs, covariance structures selection can continue much as it does for normally distributed data as long as either Laplace (preferable) or quadrature techniques are used. Moreover, while we have incorporated a parametric spatial covariance structure for the fitted negative binomial mixed-effects model, other procedures to account for spatial variation are of interest. Our study methodology, in theory, can be extended to deal with this issue using a GLMM for spatial data[29]. Therefore, we leave this and other attainable extensions for future studies.

Along this line, it would be fascinating to extend this study to the quantile mixed-effects model. Most longitudinal modeling techniques are primarily based on mean regression to focus only on the average effect of covariate and the mean trajectory of the longitudinal outcome, which is constant throughout the population. But, such average effects are not always of interest in lots of study areas and sometimes quite heterogeneous. Thus, quantile mixed-effects model has the capacity, at both the population and individual level, to discover heterogeneous covariates effects, and describe variations in longitudinal studies at different quantiles of the response variable, and hence leads to more efficient estimates, especially when the errors are over-dispersed[59,60].

Data availability

The datasets used for this study can be obtained by requesting the corresponding author on reasonable request.

Received: 28 July 2020; Accepted: 23 September 2020
Published online: 07 October 2020

References

1. WHO. HIV/AIDS: World Health Organization. Fact sheet-November (2019).
2. WHO, U. and UNICEF. Epidemiological fact sheet on HIV and AIDS core data on epidemiology and response. *South Africa, update* (2008).
3. Whelan, D. Gender and HIV/AIDS: taking stock of research and programmes (1999).
4. AMFAR. The foundation for AIDS research. 'Statistics: Women and HIV/AIDS' (2015). https://www.amfar.org/about-hiv-and-aids/facts-and-stats/statistics--women-and-hiv-aids/.
5. UN Women. Message from UN women's executive director for world AIDS day, 1 December 2014' (2014). https://www.unwomen.org/en/news/stories/2014/12/world-aids-day-2014.
6. Cohen, M. S., Shaw, G. M., McMichael, A. J. & Haynes, B. F. Acute HIV-1 infection. *N. Engl. J. Med.* **364**(20), 1943–1954 (2011).
7. Kassutto, S. & Rosenberg, E. S. Primary HIV type 1 infection. *Clin. Infect. Dis.* **38**(10), 1447–1453 (2004).
8. Rosenberg, E. S. *et al.* Immune control of HIV-1 after early treatment of acute infection. *Nature* **407**(6803), 523 (2000).
9. AIDSMAP. CD4 cell counts | aidsmap. Key points-May (2017).
10. Van Loggerenberg, F. *et al.* Establishing a cohort at high risk of HIV infection in South Africa: challenges and experiences of the CAPRISA 002 acute infection study. *PLoS ONE* **3**(4), 1 (2008).
11. Mlisana, K. *et al.* Rapid disease progression in HIV-1 subtype C-infected South African women. *Clin. Infect. Dis.* **59**(9), 1322–1331 (2014).
12. Dobson, A. J. & Barnett, A. G. *An introduction to generalized linear models* (Chapman and Hall/CRC, London, 2008).
13. Hilbe, J. M. *Negative binomial regression* (Cambridge University Press, Cambridge, 2011).
14. Hilbe, J. M. *Modeling count data* (Cambridge University Press, Cambridge, 2014).
15. Gbur, E. E., et al. Generalized linear mixed models. Analysis of Generalized Linear Mixed Models in the Agricultural and Natural Resources Sciences. American Society of Agronomy, Madison, WI**:** 109–184 (2012).
16. Stroup, W. W. *Generalized linear mixed models: modern concepts, methods and applications* (CRC Press, London, 2012).
17. Lambert, D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34**(1), 1–14 (1992).
18. Liu, W., Cela, J. Count data models in SAS. SAS Global Forum (2008).
19. Molenberghs, G. & Verbeke, G. *Models for discrete longitudinal data* (Springer, Berlin, 2006).
20. Morel, J. G. & Neerchal, N. K. *Overdispersion models in SAS* (SAS Publishing, Cary, 2012).
21. Mullahy, J. Specification and testing of some modified count data models. *J. Econ.* **33**(3), 341–365 (1986).
22. Fitzmaurice, G. M. *et al. Applied longitudinal analysis* (Wiley, Hoboken, 2012).
23. Liu, X. *Methods and applications of longitudinal data analysis* (Elsevier, New York, 2015).
24. Gill, J. & Torres, M. *Generalized linear models: a unified approach* (Sage Publications, Incorporated, 2019).
25. Guide, S. U. (2008). SAS/ETS 9.2 User's Guide, Chapter.
26. Shoukri, M. *et al.* The Poisson inverse Gaussian regression model in the analysis of clustered counts data. *J. Data Sci.* **2**(1), 17–32 (2004).
27. Wedderburn, R. W. Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. *Biometrika* **61**(3), 439–447 (1974).
28. Jiang, J. *Linear and generalized linear mixed models and their applications* (Springer, Berlin, 2007).
29. Schabenberger, O. & Gotway, C. A. *Statistical methods for spatial data analysis* (Chapman and Hall/CRC, London, 2017).

30. Lord, D., Park, B.-J., Model, P.-G. Negative binomial regression models and estimation methods. Probability density and likelihood functions. Texas A&M University, Korea Transport Institute: 1–15 (2012).
31. Demidenko, E. *Mixed models: theory and applications with R* (Wiley, Hoboken, 2013).
32. Lawless, J. F. Negative binomial and mixed Poisson regression. *Can. J. Stat.* **15**(3), 209–225 (1987).
33. Zhang, X. *et al.* Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinf.* **18**(1), 4 (2017).
34. Zhang, X. *et al.* Negative Binomial mixed models for analyzing longitudinal microbiome data. *Front. Microbiol.* **9**, 1683 (2018).
35. Fox, J. *Diagnosing problems in linear and generalized linear models: an R and S-PLUS companion to applied regression* 191–233 (Sage Publications, Thousand Oaks, 2002).
36. Allison, P. D. *Missing data* (Sage Publications, Thousand Oaks, 2001).
37. Rubin, D. B. *Multiple imputation for nonresponse in surveys* (Wiley, Hoboken, 2004).
38. Berglund, P. & Heeringa, S. G. *Multiple imputation of missing data using SAS* (SAS Institute, Cary, 2014).
39. Bucker, M. Michael J. Daniels, Joseph W. Hogan: Missing data in longitudinal studies. *Stat. Pap.* **52**(2), 501 (2011).
40. Der, G., & Everitt, B. S. Applied medical statistics using SAS. Chapman and Hall/CRC (2012).
41. Enders, C. K. Applied missing data analysis. Guilford Press (2010).
42. Fitzmaurice, G. *et al. Handbooks of modern statistical methods: Longitudinal data analysis* (Taylor & Francis Group, New York, 2009).
43. Little, R. J. & Rubin, D. B. *Statistical analysis with missing data* (Wiley, Hoboken, 2019).
44. Molenberghs, G. & Kenward, M. *Missing data in clinical studies* (Wiley, Hoboken, 2007).
45. Raghunathan, T. E. *et al.* A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodol.* **27**(1), 85–96 (2001).
46. Schafer, J. L. *Analysis of incomplete multivariate data* (Chapman and Hall/CRC, London, 1997).
47. Faraway, J. J. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models* (Chapman and Hall/CRC, London, 2016).
48. McCullaugh, P. & Nelder, J. *Generalized linear models* 2nd edn. (Chapman & Hall/CRC, London, 1989).
49. Menard, S. *Applied logistic regression analysis* (Sage, Thousand Oaks, 2002).
50. Rawlings, J. O. *et al. Applied regression analysis: a research tool* (Springer, Berlin, 2001).
51. Zeger, S. L. & Liang, K.-Y. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **1**, 121–130 (1986).
52. McArdle, B. H. & Anderson, M. J. Variance heterogeneity, transformations, and models of species abundance: a cautionary tale. *Can. J. Fish. Aquat. Sci.* **61**(7), 1294–1302 (2004).
53. Mahmud, M., et al. (2006) Selecting the optimal transformation of a continuous covariate in Cox's regression: implications for hypothesis testing. *Commun. Stat. Simul. Comput.***35**(1), 27–45.
54. McCulloch, C. E., & Neuhaus. J. M. Generalized linear mixed models. *Encyclop. Biostat.***4**, 1 (2005).
55. Shoukri, M. M. Analysis of correlated data with SAS and R. Chapman and Hall/CRC (2018).
56. Galecki, A. T. General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Commun. Stat. Theor. Methods* **23**(11), 3105–3119 (1994).
57. Harrison, X. A. *et al.* A brief introduction to mixed effects modeling and multi-model inference in ecology. *PeerJ* **6**, e4794 (2018).
58. Stroup, W. W. Rethinking the analysis of non-normal data in plant and soil science. *Agron. J.* **107**(2), 811–827 (2015).
59. Geraci, M. & Bottai, M. Linear quantile mixed models. *Stat. Comput.* **24**(3), 461–479 (2014).
60. Koenker, R. Quantile regression for longitudinal data. *J. Multivar. Anal.* **91**(1), 74–89 (2004).

Acknowledgements

We gratefully acknowledge CAPRISA for giving us access to the CAPRISA 002: Acute Infection Study data. CAPRISA is funded by the National Institute of Allergy and Infectious Diseases (NIAID), National Institutes for Health (NIH), and U.S. Department of Health and Human Services (Grant: AI51794). The authors would also like to thank Dr. Nonhlanhla Yende Zuma (Head of Biostatistics unit at CAPRISA) for her cooperation, assistance, and technical support.

Author contributions

A.A.Y. obtained the data, did the analysis, and prepared the manuscript. A.A.Y., S.F.M., H.G.M., and D.G.A. planned the research problem. All authors deliberated on the results and consequences and commented on the paper at all stages. All authors contributed extensively to the work presented in this manuscript. All authors read and ratified the ultimate manuscript.

Funding

This work was supported through the DELTAS Africa Initiative and the University of KwaZulu-Natal. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [grant 107,754/Z/15/Z], DELTAS Africa Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB) programme] and the UK government.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-73883-7.

Correspondence and requests for materials should be addressed to A.A.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

scientific reports

Check for updates

OPEN

Additive quantile mixed effects modelling with application to longitudinal CD4 count data

Ashenafi A. Yirga[1]✉, Sileshi F. Melesse[1], Henry G. Mwambi[1] & Dawit G. Ayele[2]

Quantile regression offers an invaluable tool to discern effects that would be missed by other conventional regression models, which are solely based on modeling conditional mean. Quantile regression for mixed-effects models has become practical for longitudinal data analysis due to the recent computational advances and the ready availability of efficient linear programming algorithms. Recently, quantile regression has also been extended to additive mixed-effects models, providing an efficient and flexible framework for nonparametric as well as parametric longitudinal forms of data analysis focused on features of the outcome beyond its central tendency. This study applies the additive quantile mixed model to analyze the longitudinal CD4 count of HIV-infected patients enrolled in a follow-up study at the Centre of the AIDS Programme of Research in South Africa. The objective of the study is to justify how the procedure developed can obtain robust nonlinear and linear effects at different conditional distribution locations. With respect to time and baseline BMI effect, the study shows a significant nonlinear effect on CD4 count across all fitted quantiles. Furthermore, across all fitted quantiles, the effect of the parametric covariates of baseline viral load, place of residence, and the number of sexual partners was found to be major significant factors on the progression of patients' CD4 count who had been initiated on the Highly Active Antiretroviral Therapy study.

Abbreviations

AMM	Additive mixed model
QR	Quantile regression
AQM	Additive quantile model
AQMM	Additive quantile mixed model
GAMLSS	Generalized additive model for location, scale, and shape
CAPRISA	Centre of the AIDS Programme of Research in South Africa
HIV	Human immunodeficiency virus
AIDS	Acquired immune deficiency syndrome
CD4	Cluster of difference 4 cell (t-lymphocyte cell)
VL	Viral load refers to the number of HIV copies in a milliliter of blood (copies/ml)
STD	Sexually transmitted diseases
ART	Antiretroviral therapy
ARV	Antiretroviral (drug)
HAART	Highly active antiretroviral therapy
WHO	World Health Organization

Parametric models relate the mean of a response variable to a linear combination of covariate effects and focus on the response's average properties[1]. Nevertheless, there are inevitable occasions when such parametric models fail, and data analysis must turn to more flexible, nonparametric models[2]. Parametric models also assume a distribution for the outcome variable as opposed to purely nonparametric models. However, most of the vast literature on nonparametric regression also deals with the estimation of conditional mean models. In addition, the conventional assumption of nonparametric regression theory that there is additive, independently, and identically distributed (*iid*) error around a smooth underlying conditional mean function is highly implausible in certain data settings[2]. Thus, as in the parametric context, nonparametric methods are usefully complemented

[1]School of Mathematics, Statistics, and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, Private Bag X01, Scottsville 3209, South Africa. [2]Institute of Human Virology, School of Medicine, University of Maryland, Baltimore, MD 21201, USA. ✉email: ashu3argaw@gmail.com

by nonlinear estimation of families of conditional quantile functions that relax the independence assumption[2]. The use of parametric and nonparametric regression models for analyzing patients' CD4 count in most applications implies that the estimated effects describe the average CD4 count. However, it is of even great interest to examine the quantile of the outcome distribution, such as the lower ($\leq 25\%$) quantile, which identifies patients at higher risk of developing illnesses.

Quantiles, commonly symbolized by the Greek letter τ, are location and scale parameters simultaneously. For a given $\tau \in (0, 1)$, the τ^{th} quantile is the value of a random variable, where $\tau \times 100\%$ of its value lies below it. In other words, it is the value where at most $(1 - \tau) \times 100\%$ of the value lies above. Thus, τth quantiles close to 0.5-quantile give the median, which is a well-known location parameter. On the other hand, τth quantiles close to zero or one give an idea of the scale. For instance, the interquartile range (IQR) is defined as the 0.75 quantile minus the 0.25 quantile: $IQR = Q_3 - Q_1$.

Quantile regression (QR) solutions are computed for a selected number of quantiles, typically the three quantiles along with two extreme quantiles, that is, for $\tau = \{0.05, 0.25(Q_1), 0.5(Q_2), 0.75(Q_3), 0.95\}$. This necessitates the search for a suitable compromise between the amount of output to manage and the results to interpret and summarize. Although in many practical applications of QR, the focus is on estimating a subset of quantiles, however, it is worth noticing that it is possible to attain estimates across the entire interval of conditional quantiles; in particular, the set: $\{\beta_\tau : \tau \in (0, 1)\}$[2].

QR is a versatile statistical method with many applications that complement mean regression[3,4]. Thus, it emerged as an effective analytic technique in numerous study areas of science due to its competence to drive inferences about individuals that rank below or above the conditional population mean and/or focused on features of the response beyond its central tendency[4–13]. QR is specifically appropriate for the parameters' heterogeneous effect as it yields inferences that can be legitimate irrespective of the true underlying distribution[4,14]. QR techniques look further into the data, get more information, and become more important[15]. By fitting models for more percentiles, one can detect the covariates' heterogeneous effects at the conditional distribution of the response, rather than just the conditional mean. That is especially useful when valuable information lies at the bottom or top quantiles. "QR also enjoys several properties, including equivariance to monotone transformations and robustness to outliers"[2,16]. A semiparametric extension of quantile regression models with different types of nonlinear effects included in the model equation leads to an additive quantile regression model (AQM)[12]. Such a model may reveal systematic differences in dispersion, tail behavior, and other features for covariates[2].

Additive mixed models (AMMs), an extension of additive models, have been developed precisely to incorporate linear and nonlinear effects, as well as random terms when the data are sampled according to longitudinal designs[4,17]. AMMs have been integrated into QR methods to obtain robust results, not only focused on features of the longitudinal outcome at its central tendency that may not be the best location to characterize the data specifically when the errors are non-normally distributed, and the location-shift hypothesis of the normal model is violated but also at conditional quantiles of the longitudinal outcome with no assumption about the response or errors distribution apart from the distribution is restricted to have the τth quantile to be zero. Thus, additive quantile mixed models, which have gained popularity recently as a general method for longitudinal data, bring a comprehensive and more complete picture of the nonparametric as well as the parametric effects[1,4].

CD4 cell count levels signify the well-being of an individual immune system (body's natural defense system against pathogens, infections, and illnesses). The CD4 cell counts of a person who does not have HIV can be between 500 and 1500 per cubic millimeter. Individuals living with HIV who have a CD4 count over 500 but whose immune response is still strong are usually in good health. However, individuals living with HIV who have a CD4 count below 200 are at high risk of developing severe illnesses and death[18,19].

With the CD4 count at deficient levels, patients' immunity is weak. If HIV-infected patients are not on treatment or not virally suppressed, they become vulnerable to acquire opportunistic infections (OIs), making them at risk of the new and ongoing coronavirus disease 2019 (COVID-19) infection and underlying illness[18]. The best strategy to avoid these infections and diseases is by enhancing the immune function level through HAART, a combination of multiple antiretroviral (ARV) drugs. HAART's fundamental goal is to prolong or stop the progression to AIDS and loss of life for those infected with HIV by suppressing and preventing the virus from making copies of itself. When the virus's level (viral load) in the blood is low or undetectable, there is less damage to the body's immune system and fewer HIV infection complications. Even though HIV treatment is prescribed for all individuals living with HIV, it is particularly critical for patients with low CD4 count to start treatment sooner rather than later and adhere to the treatment schedule[18,20]. While researchers believe that early diagnosis and effective treatment are essential to effective control, more research is needed to understand better the adaptive, innate, and host responses that alter viral load set-point and consequently prognosis and infectiousness[18,20].

The need for good and better health is one of each human being's fundamental rights without qualification of race, religion, gender, political conviction, financial, or social condition. Women's health includes their emotional, social, and physical welfare and is determined by these factors and the economic setting of their lives, as well as by biology. However, health issues evade the longer part of women. In national and universal forums, women have emphasized that equality, the sharing of family duties, development, and peace are necessary conditions to achieve good health all through the life cycle. Women are biologically and socially more vulnerable to HIV infection, especially in developing countries[21–24].

HIV/AIDS and other sexually transmitted diseases (STD) have a devastating effect on women's health, mostly young ladies. The consequences of HIV/AIDS go beyond women's health to include their families' economic support and livelihoods. Thus, the social, development, and health consequences of HIV/AIDS and other sexually transmitted diseases have strong gender dimensions that cannot be ignored[23–25]. Understanding the changing epidemiology of HIV using statistical disease models will allow the clinician to decide who may be at high risk and clarify the application of rules to avoid sequential HIV transmission[18,20,26,27]. Although antiretroviral (ARV) recommendation presently remains the same for all individuals living with HIV, examining the progression of

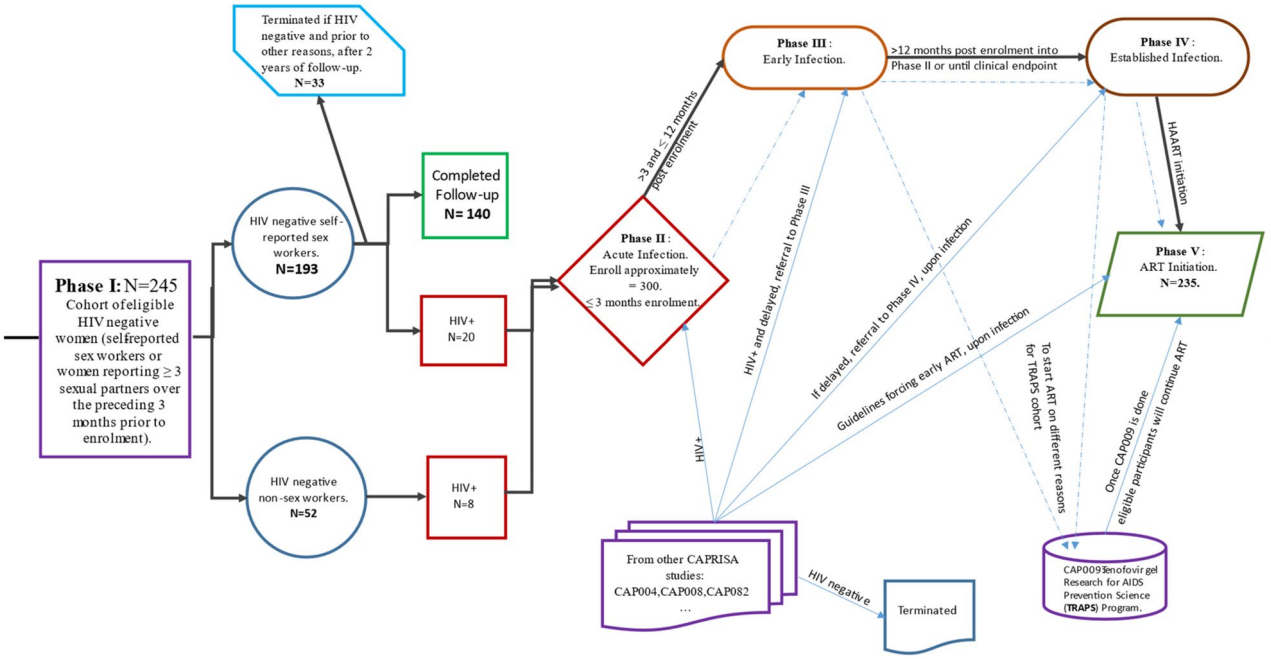


Figure 1. Diagrammatic overview of the CAPRISA 002 AI cohort study design.

CD4 count or evolution of the viral load using data-driven models will allow the clinician to interpret potential information accurately and cope with misdirection or distortion of the information due to patient-specific effects[18,26–28]. This study is a continuation of our previous work in Yirga et al.[18]. This study aims to analyze the longitudinal CD4 count of HIV-infected patients involved in a CAPRISA study using AQMM and justify how the method evolved can be used to attain robust nonparametric as well as parametric effects at various locations of the conditional distribution that brings a comprehensive and more complete picture of the covariate effects. The use of AQMM has many advantages. Additive nonparametric effects models are not new in the applied statistics literature. To implement these methods, Koenker et al.[47] introduce smoothing penalties for total variation, especially for the nonparametric components of the model. Researchers are also eager to learn what are the factors influencing the CD4 count (high or low) in HIV studies. AQMMs are the best way to answer this question.

Materials and methods

Data description. This study used data from the Centre for the AIDS Programme of Research in South Africa (CAPRISA). The CAPRISA study was effected at the Doris Duke Medical Research Institute (DDMRI) at the Nelson R Mandela School of Medicine of the University of KwaZulu-Natal in Durban, South Africa[18,29]. Between August 2004 and May 2005, CAPRISA introduced a cohort study registering high-risk HIV-negative women to a follow-up study with an intense ongoing examination. Women infected with HIV were recruited into the CAPRISA 002 Acute Infection (AI) study and then followed up carefully to study disease progression and CD4/viral load evolution[18,20,29–32].

Once HIV-infected women were enrolled in CAPRISA's AI Phase II study, their CD4 count and viral load were measured and assessed regularly. When their CD4 count $\leq$ 350 cells/mm^3 for more than two consecutive visits between six months or if they are with AIDS-defining illness (WHO clinical stage 3–5), they would be referred to a public government clinic for ARV treatment. However, according to the South African National Department of Health, these patients would only start HAART once their CD4 count is $\leq$ 200 cells/mm^3, until 2015. With effect from the 1st of January 2015, according to the National Department of Health, the criteria to start HIV patients on early initiation of ART is CD4 count of 500 cells/mm^3 or less than that[20]. HIV-infected women in Phase II–IV were followed up until they are started HAART. After that, they would be transitioned to Phase V and followed up for a minimum of five years, or eligible participants would be offered to join immediately into Phase V[33]. After the five years of follow-up have been accomplished, participants would be offered an optional annual follow-up for up to fifteen extra years to patients who recurred in Phase V[33]. Figure 1 illustrates the screening and enrolment process of the study data set. One can find further detail on the study population's design, development, and procedures here[29–33].

Consent for publication. Not applicable.

Methods

Parametric regression models typically use a linear function to connect the conditional values of the response variable to the covariates. In real-world applications, however, biased or invalid results might result from such a linearity assumption. Many studies use nonlinear assumptions between variables[34–37]. One may consider various

modeling techniques when dealing with nonlinearity. The most popular nonparametric models, smoothing splines, and transformation models use parameters such as sampling designs (cross-sectional or longitudinal), outcomes (discrete or continuous), distribution assumptions (parametric or nonparametric), and so on[2]. In choosing which method to follow, the amount of effort expended during the investigation may have a significant influence. Likewise, lacking theory or programming can lead to a certain decision being made over another[2].

Nonparametric regression permits the presumption of linearity to be relaxed[34,35,38] and limits the analysis to smooth and continuous functions[39]. Nonparametric regression, also known as scatter smoothing, aims to distinguish the best regression function according to the data distribution instead of estimating the parameters[39].

The nonparametric regression model is given by.

$$y = \sum_{i=1}^{n} f_i(x_i) + \varepsilon_i, \tag{1}$$

where the function $f_i(\cdot)$ is unknown, and commonly assumed that the errors are normally and identically distributed: $\varepsilon_i \sim NID\left(0, \sigma^2\right)$[39]. Several methods have been introduced to model nonparametric regression models; however, the most used techniques that have been extended to QR are local polynomial regression[40] and smoothing splines[41,42]: for further details, see Wu and Zhang[34], Fox[38], Davino et al.[39], Craig and Ng[43], Koenker et al.[44], Koenker[45], Cleveland and Loader[46], or Koenker et al.[47].

The parametric QR model is given by.

$$Y_i = \boldsymbol{x}_i^{'}\boldsymbol{\beta}_{\tau i} + \varepsilon_{\tau i}, \quad i = 1, \ldots, n, \quad 0 < \tau < 1, \tag{2}$$

where Y_i is the response variable, x_i's are covariates, $\beta_{\tau i}$'s are the quantile specific linear effects, and $\varepsilon_{\tau i}$ is a random variable assumed to be an unknown error term on which no specific distributional assumptions are made except that the distribution is restricted to have the τth quantile to be zero[12,48,49]. For this reason, the parametric QR model aims at describing the quantile function $Q_{Y_i}(\tau|\boldsymbol{x}_i)$ of the continuous outcome Y_i conditional on covariate vector $\boldsymbol{x}_i$ at a given quantile τ, and this can be expressed as follows

$$Q_{Y_i}(\tau|\boldsymbol{x}_i) = F_{Y_i}^{-1}(\tau|\boldsymbol{x}_i) = \boldsymbol{x}_i^{'}\boldsymbol{\beta}_{\tau i} + \varepsilon_{\tau i}, \quad \text{with } Q_{\varepsilon_{\tau i}}(\tau|\boldsymbol{x}_i) \sim F_{\tau i}, \tag{3}$$

where $F_{\tau i}$ is subject to $F_{\tau i}(0) = \tau$, $F_{Y_i}^{-1}(\cdot)$ is the inverse cumulative distribution function of Y_i. For a comprehensive overview of QR, see, for example, Koenker[2], Konker and Basset[3], Buchinsky[5], Yu et al.[9], or Koenker and Hallock[50].

As much as the parametric QR assumptions enjoy a simple model structure, convenience of interpretation, and lower computational cost, it is not flexible enough and hence carries the risk of model misidentifications for complex problems[51]. Nonparametric QR has become a viable alternative to avoid restrictive parametric assumptions. Koenker et al.[47] explored nonparametric QR in spline models (quantile smoothing splines), which they defined as solutions to

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \rho_\tau\left(y_i - f(x_i)\right) + \lambda \left(\int_0^1 |f''(x)|^p dx\right)^{1/p}, \tag{4}$$

where $\rho_\tau(u) = u\{\tau - I(u < 0)\}, p \geq 1$, is the so-called *check (loss) function,* the parameter $\tau \in (0, 1)$ controls the quantile of interest, and $\lambda \in \mathbb{R}^+$ is a smoothing parameter[3,47].

As closely analogous to the parametric QR model (3), Koenker[2] generalized nonparametric QR models as

$$Q_{Y_i}(\tau|\boldsymbol{x}_i) = f(\boldsymbol{x}_i, \beta_i(\tau)) \tag{5}$$

Then, Koenker[2] formulated the τth nonparametric QR estimator as

$$\hat{\boldsymbol{\beta}}_i(\tau) = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau\left(y_i - f(\boldsymbol{x}_i, \beta(\tau))\right) \tag{6}$$

Several techniques were proposed for nonparametric QR modelings, such as Bivariate quantile smoothing spline[52] and Kernel quantile regression[53]. However, nonparametric QR is an important yet challenging topic that needs to be addressed in-depth[51]. One can find a brief account of nonparametric QR strategies in numerous studies; see, for example, Koenker[2] or Davino et al.[39]. To account for the nonlinearity relationships between quantiles of the outcome and covariates, Rigby and Stasinopoulos[54] also proposed generalized additive models for location, scale, and shape (GAMLSS). GAMLSS enables additional flexibility to fit the covariates' nonlinear effects; however, they do not result in easily interpretable expressions for the quantiles. They are based on specifying distinct distributional parameters[12]. Instead, additive quantile regression models (AQMs) allow for the inclusion of nonlinear covariate effects and give more flexibility[12].

Additive models, introduced by Hastie and Tibshirani[41], Stone[55], and Breiman and Friedman[56], are flexible regression tools that manipulate linear as well as nonlinear terms. The nonlinear terms in additive models are modeled through smoothing splines[4]. They provide programmatic approaches for nonparametric (nonlinear in parameters) regression modelings; by restricting nonlinear covariate effects to be composed of low-dimensional additive pieces so that we can overcome some of the worst aspects of the notorious curse of dimensionality[11]. The literature on additive models is vast[17,41,55,57,58]. However, most of the work has been done based on estimating conditional mean functions. The additive quantile regression model (AQM) provides an attractive framework

for parametric as well as nonparametric regression illustrations focused on features of the response beyond its central tendency[4,11,12].

Fenske et al.[12] defined the τth AQMs that extend the linear predictor, $\boldsymbol{x}_i^{'}\boldsymbol{\beta}_\tau$, with a sum of nonlinear functions of continuous covariates, $\sum f_{\tau j}(\cdot)$, as follows.

$$Q_{Y_i}(\tau|\boldsymbol{x}_i,\boldsymbol{z}_i) = \boldsymbol{x}_i^{'}\boldsymbol{\beta}_{\tau i} + \sum_{j=1}^{q} f_{\tau j}(\boldsymbol{z}_i) + \varepsilon_{\tau i}, \quad j = 1,\ldots,q, \tag{7}$$

where $f_{\tau j}$ denote generic functions of covariates $\boldsymbol{z}_i$ for the ith observation, and allows for the inclusion of different model terms such as nonlinear effects (smooth functions) of z_k, $f_\tau(z_k)$, and varying coefficient terms, $z_k^{'}f_\tau(z_k)$, where the effect of the covariate $z_k^{'}$ varies smoothly over the domain of z_k according to some functions of f_τ. However, the underlying assumption of the error term, $\varepsilon_{\tau i}$, remains the same as in the QR model (3); see Fenske et al.[12] for more details.

AQM estimates the additive effect using linear programming algorithms as in the conventional QR model[12]. However, in the AQM case, determining adequate numbers and the position of knots is challenging. To avoid these challenges, Fenske et al.[12] used penalty methods such as quantile smoothing splines of Koenker et al.[47]. Thus, the minimization problem of AQM that consists of extra penalty term is given by[12]:

$$\underset{f_\tau}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau \left(y_i - \boldsymbol{x}_i^{'}\boldsymbol{\beta}_{\tau i} - \sum_{j=1}^{q} f_{\tau j}(\boldsymbol{z}_i) \right) - \lambda \boldsymbol{v}\left(f_\tau^{'}\right), \tag{8}$$

where $\boldsymbol{v}\left(f_\tau^{'}\right) = sup \sum_{i=1}^{n-1} \left| f_\tau^{'}(z_{i+1}) - f_\tau^{'}(z_i) \right|$, represents the total variation of the derivation $f_\tau^{'} : [a,b] \to \mathbb{R}$, where the *sup* is taken over all partitions $a \le z_1 < \ldots < z_n < b$, and λ is a tuning parameter that controls the smoothness of the estimated function also known as "total variation regularization": see Koenker[2], Fenske et al.[12], or Koenker et al.[47] for more details.

Fenske et al.[1] proposed extending AMMs to the QR model for longitudinal data that consists of fixed individual-specific intercepts and slopes modeled through penalized splines of Ruppert et al.[59]. However, their model did not include random-effect terms and did not allow for individual-specific effects to have a general covariance structure[4]. The version of Geraci[4] additive QR model for longitudinal data includes linear and nonlinear terms, as well as multiple random effects to account for the correlation at the individual level with a general variance–covariance matrix and allow for automatic smoothing selection within a mixed model framework of Ruppert et al.[59]. Thus, as pointed out by Geraci[4], because of the following two basic ideas, his model was shown to have superior performance compared with the approach of Fenske et al.[1]: the first point is regarding the ith unit effects, which he assumed to be random instead of fixed so that the covariance structure between effects can be introduced; the second point is that instead of prior specification, the nonparametric term's smoothing is automatically estimated from the data[4].

Geraci[4] defined the τth additive QR model for longitudinal data as

$$Q_{y_{ij}|\boldsymbol{u}_i,\boldsymbol{x}_i,\boldsymbol{z}_i}(\tau) = \beta_{\tau,0} + \sum_{k=1}^{p} f_\tau^k\left(x_{ijk}\right) + z_{ij}^{'}u_{\tau,i}, \tag{9}$$

$$j = 1,\ldots,n_i, \quad i = 1,\ldots,m, \quad \tau \in (0,1),$$

where $x_{ij}^{'}$ is the jth row of a known $n_i \times p$ matrix $\boldsymbol{X}_i$, $z_{ij}^{'}$ is the jth row of a known $n_i \times q$ matrix $\boldsymbol{Z}_i$, y_{ij} is the jth observation of the response vector $\boldsymbol{y}_i = \left(y_{11},\ldots,y_{1n_i}\right)$ for the ith unit, $f_\tau^k(\cdot)$ is a τ-specific, centered, twice-differentiable smooth function of the kth component of $\boldsymbol{x}$, and $u_{\tau,i}$ is a $q \times 1$ vector of values that collects ith unit random effects associated with z_{ij} and its distribution is assumed to depend on a τ-specific parameter[4].

Geraci[4] considered a spline model of the type: $f_\tau(x) \approx \sum_{h=1}^{H} \nu_{\tau,h} B_h(x)$, to model nonlinear functions of the components of $\boldsymbol{x} = (x_1,\ldots,x_s,x_{s+1},\ldots,x_p)$ that consists of the first s terms of nonlinear functions and $p-s$ linear functions. The B_h's denote the *basis functions* (ν_τ), h's represent the corresponding τ-specific coefficients of B_h's and H indicates the number of knots[4]. The approximated quantile function from the model (9) is then expressed as follows[4]:

$$Q^*_{y_{ij}|\boldsymbol{u}_i,\boldsymbol{x}_i,\boldsymbol{z}_i}(\tau) = \beta_{\tau,0} + \sum_{k=1}^{s}\sum_{h=1}^{H_k} \nu_{\tau,hk} B_h^{(k)}\left(x_{ijk}\right) + \sum_{k=s+1}^{p} \beta_{\tau,k}\, x_{ijk} + z_{ij}^{'}u_{\tau,i} \tag{10}$$

In matrix notation, the ith unit of expression (10), which is then called additive quantile mixed model (AQMM), is given by[4]

$$Q^*_{\boldsymbol{y}_i|\boldsymbol{u}_i,\boldsymbol{x}_i,\boldsymbol{z}_i}(\tau) = \boldsymbol{F}_i\boldsymbol{\beta}_\tau + \boldsymbol{Z}_i\boldsymbol{u}_{\tau,i} + \boldsymbol{B}_i\boldsymbol{\nu}_\tau, \tag{11}$$

where $B^{(k)}\left(x_{ijk}\right)$ is considered as $H_{k\times 1}$ vector of values taken by the kth spline evaluated at x_{ijk}, $\nu_{\tau,k} = \left(\nu_{\tau,1},\ldots,\nu_{\tau,H_k}\right)^{'}$ considered as the $H_{k\times 1}$ vector of spline coefficients for the kth covariate, and $H = \sum_k H_k$. Furthermore, $\boldsymbol{B}_i$ and $\boldsymbol{\nu}_\tau$, defined, respectively, as the $n_i \times H$ matrix with rows $\left(B^{(1)}\left(x_{ij1}\right)^{'},\ldots,B^{(s)}\left(x_{ijs}\right)^{'}\right)$ and

Variable	Descriptive measures						
	Mean	Median	Minimum	Maximum	$Q_{0.25}$	$Q_{0.75}$	IQR
SQRT_CD4 count (cells/μL)	23.26	22.98	5	44	20	26.19	6.19
Baseline VL (cells/**mL**)	130,730.33	26,600	1 (undetected)	5,510,000	5080	113,000	107,920
Age (Years)	27.15	25	18	59	22	30	8
Baseline BMI	28.98	26.84	17.89	54.89	23.33	32.96	9.63

Table 1. Descriptive statistics for non-categorical variables.

Variable	Total	Variable	Total
Place of residence		**Number of sexual partners**	
Rural (*reference*)	105 (44.7%)	No partner (*reference*)	43 (18.3%)
Urban	130 (55.3%)	Stable partner	182 (77.4%)
Educational level		Many partners	10 (4.3%)
Primary schools (*reference*)	11 (4.7%)	**Number of women**	235
Secondary schools	224 (95.3%)		

Table 2. Baseline descriptive statistics for categorical variables.

$\left(v'_{\tau,1},\ldots,v'_{\tau,s}\right)'$, $\boldsymbol{F}_i$ is the $n_i \times (p-s+1)$ matrix with rows $(1, x_{ij(s+1)},\ldots,x_{ijp})'$ and $\boldsymbol{\beta}_\tau = (\beta_{\tau,0}, \beta_{\tau,s+1},\ldots,\beta_{\tau,p})'$[4].

The objective function of AQMM, where the vectors $\boldsymbol{u}_{\tau,i}$ and $\boldsymbol{v}_\tau$ are assumed to follow zero-centered multivariate Gaussian distributions with variance–covariance matrices $\boldsymbol{\Sigma}_\tau$ and $\boldsymbol{\Phi}_\tau = \oplus_{k=1}^{s}\phi_{\tau,k} I_{H_k}$, respectively, with selecting $\rho_\tau(\boldsymbol{r}) = \sum_{j=1}^{n} r_j\{\tau - I(r_j < 0)\}$ for a vector $\boldsymbol{r}=(r_1,\ldots,r_n)'$, is given by Geraci[4] as

$$\sum_{i=1}^{M}\rho_\tau\left(\boldsymbol{y}_i - \boldsymbol{F}_i\boldsymbol{\beta}_\tau - \boldsymbol{Z}_i\boldsymbol{u}_{\tau,i} - \boldsymbol{B}_i\boldsymbol{v}_\tau\right) + \sum_{i=1}^{M}||u_{\tau,i}||^2_{\Sigma_\tau^{-1}} + \sum_{k=1}^{s}\phi_{\tau,k}^{-1}||v_{\tau,k}||^2, \tag{12}$$

where "$\boldsymbol{u}_{\tau,i}$'s are assumed to be independent for different i (but may have a general covariance matrix) and are independent of $\boldsymbol{v}_\tau$, and $\phi_{\tau,k}$'s determine the amount of smoothing for the nonparametric terms"[4]. Minimizing the objective function of expression (12) proceeds as the same as minimizing the objective function of quantile mixed-effects models[49,60,61] where the asymmetric Laplace distribution with a location parameter μ, scale parameter $\sigma > 0$, and skewness parameter $\tau \in (0,1)$[60,62–64], are employed as *quasi-likelihood* for the fidelity term[4]. Further discussion of AQMM is provided by Geraci[4].

Ethical approval and consent to participate. The study was approved by the Research Ethics Committee of the University of KwaZulu-Natal (E013/04), the University of the Witwatersrand (MM040202), and the University of Cape Town (025/2004). All participants provided written informed consent. All methods were performed following the relevant guidelines and regulations expressed in the Declaration of Helsinki.

Results

Geraci[4] illustrated the full range of AQMM that is described above. The purpose of this analysis is to model the CD4 count of patients from KwaZulu-Natal, South Africa, as part of a comprehensive study of HIV/AIDS. The results of this study illustrate longitudinal CD4 counts among HIV-infected patients enrolled in the CAPRISA 002 AI study by employing an AQMM. The median age of our sample of 235 women was 25 years. Our sample consisted of 7019 measurements on 235 women from 18 to 59 years of age. There were multiple visits for all participants, ranging from 2 to 61, with a median of 29.

Tables 1 and 2 show descriptive measures for the variables studied. Low (upper) quantiles are those where at least 25% (75%) of the observations are at or below it, or 75% (25%) are at or above it[2]. In Table 1, it is shown that the median BMI for the participants was 26.84 (range 17.89–54.89). The median square root CD4 count and baseline viral load were 22.98 cells/mm^3 and 26,600 copies, respectively. Of a total of 235 women, 105 (44.7%) lived around Vulindlela (rural area), and 130 (55.3%) lived around eThekwini (Durban, urban area) in KwaZulu-Natal, South Africa (see Table 2). The majority of the women, 182 (77.4%), were in a stable partnership, 224 (95.3%) completed secondary school (Table 2), and most of them (78.8%) were self-reported sex workers[18,29,31]. Additional details are available here[29–32] concerning the CAPRISA 002 AI study. We analyze this study data set intending to explain the different conditional distribution of the CD4 count by considering two covariates entered as nonparametric additive effects: time and baseline BMI; as well as discrete (baseline viral load), continuous (age), and categorical covariates (place of residence, educational level, and the number of sexual partners) entered in the model as parametric effects (see Tables 1, 2). Figure 2 shows observed square root transformed CD4 counts

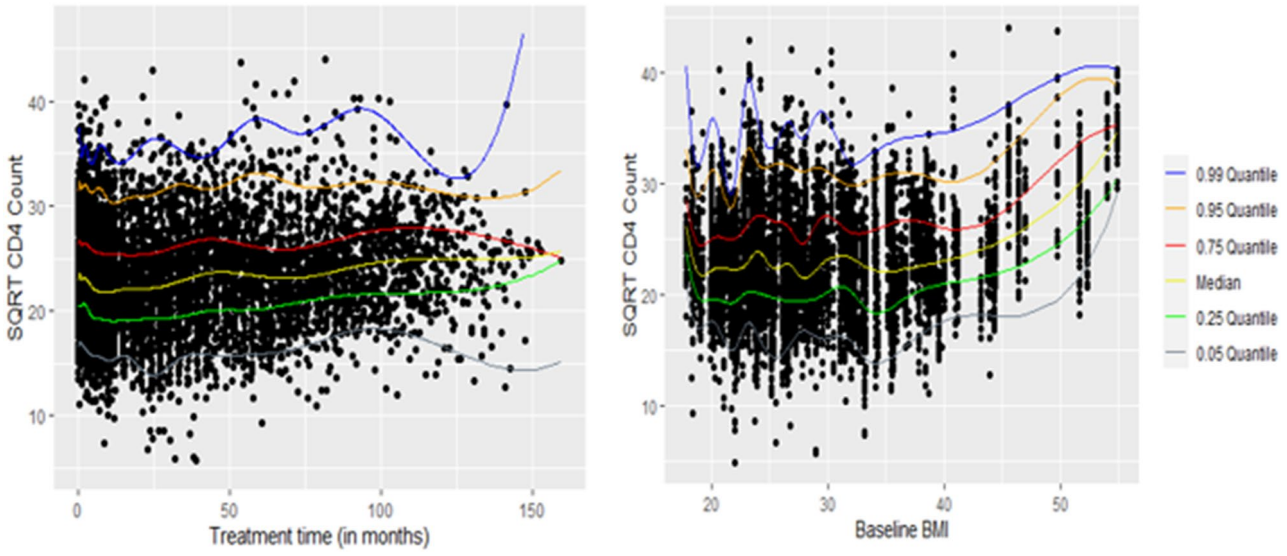


Figure 2. Observed CD4 counts (square root transformed) by time and baseline BMI.

by treatment time and baseline BMI, respectively, for a total of 7019 observations. The nonlinear patterns, which connect the sample quantiles, are estimated conditionally on time and baseline BMI for six quantile levels. The curves (nonlinear patterns) suggest the requirement of some degree of smoothing (Fig. 2).

Following the AQMM of Geraci[4], we used a transformed continuous form of the outcome (i.e., square root CD4 count) for fitting purposes. Thus, the proposed τ^{th} AQMM form of our study, using expression (10), can be specified as

$$\begin{aligned} Q^{*}_{y_{ij}|\boldsymbol{u}_i,\boldsymbol{x}_i,\boldsymbol{z}_i}(\tau) = \beta_{\tau,0} + \sum_{h=1}^{H_1} \nu_{\tau,1}\, B_h^{(1)}(time_i) + \sum_{h=1}^{H_2} \nu_{\tau,2}\, B_h^{(2)}(BMI_i) + \beta_{\tau,1}\, ART_i \\ + \beta_{\tau,2}\, VL_i + \beta_{\tau,3}\, residence_i + \beta_{\tau,4}\, education_i + \beta_{\tau,5}\, partner_i \\ + \beta_{\tau,6}\, age_i + u_{\tau,0} + u_{\tau,1}(time_i), \end{aligned} \tag{13}$$

where y_{ij} is the square root transformed form of the outcome ($\sqrt{CD4count}$) at the jth time point for the ith subject, time is the time variable measured in months from the start of the study, BMI indicates the patient's baseline BMI, ART is the dichotomous HAART initiation (0 = pre-ART, 1 = post-ART), VL is patient's baseline viral load, the residence is patient's place of residence, education is the educational level of participants, partner indicates the number of sexual partners of the participant, age is participant age at enrolment, $u_{\tau,0}$ indicates the random intercept, and $u_{\tau,1}$ indicates the random slope. The symbol τ specifies the quantile of interest; we made the estimation at $\tau = 0.05, 0.25, 0.5, 0.75, 0.85, 0.95$, and 0.99 to get the complete picture of the effects.

Geraci[4] employed the AQMM in the R package *lqmm* as an ad-on to fit additive quantile mixed models. As the same as the smooth terms' specification in the R package *mgcv*[17], one can enter continuous covariates within the *s* (smooth) function to control the model smoothness using splines when fitting AQMM[4]. Furthermore, the shrinkage smoothers obtained using the *bs* option inside the *s* command in the R package *mgcv* are constructed so that smooth terms can be penalized away altogether, not contribute to the model[17,65]. Thin plate smoother provides statistical and computational efficiency, stable optimal approximations (especially for large data sets), and can be constructed for smooths of more than one covariate at a time[4,66]. Thus, it was used as a shrinkage spline to fit the proposed model (13). The remaining parametric terms in the *aqmm* function[4] are specified the same way as in other R linear mixed model fitting functions such as *lqmm ()* and *lme4 ()*. The output is separated into two parts: Parametric part that includes estimated fixed effects, with their standard errors (SE), in parentheses, and significant mixed effect representation of smoothing splines (see Table 3). Since the smooth coefficients are mostly uninterpretable, we focus on their variances to evaluate the spline coefficients' penalty at various quantiles (see Table 4 and Supplementary information). However, their estimated smoothed effects are depicted in Fig. 3. Table 4 also presents the estimated variance of the random effects from the fitted model (13).

According to Table 3, the age effect is positive and significant at the bottom, median, and at $\tau = 0.75$ quantile levels (see also Supplementary information). On the other hand, the effect of education on square root CD4 count does not seems to be significant across all quantiles after the patient had been initiated on HAART. The square root CD4 count across all quantiles is affected by post-HAART initiation as expected. A significant positive effect of HAART initiation on CD4 cell counts is observed at the median quantile and becomes roughly constant at higher quantiles (see Table 3 and Supplementary information). In addition, patients with stable sexual partners showed significant improvements in their CD4 cell count across all quantiles. The CD4 cell count is significantly lowered in patients who have many sexual partners, especially at the bottom ($\tau = 0.05$) and at the top ($\tau = 0.95, 0.99$) quantiles (Table 3).

Fixed effects	$\widehat{Q}_{0.05}$ (SE)	$\widehat{Q}_{0.25}$ (SE)	$\widehat{Q}_{0.5}$ (SE)	$\widehat{Q}_{0.75}$ (SE)	$\widehat{Q}_{0.95}$ (SE)
Intercept	16.004 (0.6634)***	19.647 (0.4749)***	21.204 (0.5340)***	24.167 (1.0536)***	29.379 (0.6324) ***
Age	0.0398 (0.0156)**	0.0209 (0.0114)	0.0418 (0.0052)***	0.0331 (0.0078)***	0.0203 (0.0178)
Secondary school	− 0.4491 (0.5731)	− 0.4734 (0.4101)	− 0.0165 (0.6619)	0.0385 (1.0677)	0.8323 (0.5574)
Post HAART	0.7430 (0.0879)***	1.5296 (0.0598)***	1.5968 (0.0402)***	1.5292 (0.0546)***	1.7007 (0.1322)***
Baseline VL	− 3.83e−06 (8.42e−07)***	− 2.09e−06 (2.69e−07)***	− 1.79e−06 (2.41e−07)***	− 1.57e−06 (1.60e−07)***	− 1.70e−06 (2.21e−07)***
Urban	− 0.50002 (0.1668)**	0.2499 (0.0545)***	0.0998 (0.0334)**	0.1275 (0.1436)	− 0.8846 (0.2216)***
Stable partner	0.6135 (0.1655)***	0.3046 (0.1549)	0.5424 (0.1140)***	0.4907 (0.1594)**	0.6339 (0.2960)*
Many partners	− 2.2771 (0.2707)***	− 0.7858 (0.2589)**	− 0.8432 (0.1091)***	− 1.1719 (0.2569)***	− 3.6497 (0.4451)***
Results of the smooth terms					
s (Time)	− 2.5075 (0.5426)***	− 2.3766 (0.5549)***	− 2.1985 (0.4735)***	− 2.2829 (0.4999)***	− 2.3324 (0.4373)***
s (Baseline BMI)	5.4382 (1.0786)***	5.6868 (1.1094)***	5.5767 (1.3014)***	5.7904 (1.2077)***	5.2604 (1.0753)***

Table 3. Parameter estimates followed by results of the smoothing terms from the AQMM for the CAPRISA 002 AI study data across different quantiles. *Significance codes: 0 '***', 0.001 '**', 0.01 '*', 0.05 '', 0.1 ' ', 1. The reference categories are given in Table 2.

Results across different quantiles							
	$\hat{Q}_{0.05}$	$\hat{Q}_{0.25}$	$\hat{Q}_{0.5}$	$\hat{Q}_{0.75}$	$\hat{Q}_{0.85}$	$\hat{Q}_{0.95}$	$\hat{Q}_{0.99}$
Variance of the random effects							
$\hat{\sigma}_0$(Intercept)	0.02748	0.8687	0.0354	0.2453	0.3454	0.0467	0.0033
$\hat{\sigma}_0$(Time)	8.104e−18	1.929e−16	3.328e−17	5.451e−17	7.671e−17	1.044e−17	2.963e−18
Variance of the smooth terms							
$\hat{\phi}_{Time}$	8.796	28.94	36.74	30.28	21.92	10.13	2.669
$\hat{\phi}_{BaselineBMI}$	1876.501	6463.83	7823.81	6290.32	4979.39	2183.69	576.902

Table 4. Estimated variance of the random effects and smooth terms from the AQMM for the CAPRISA 002 AI study data.

Furthermore, we found a clear indication, at the bottom ($\tau = 0.05$) and more extreme quantiles ($\tau = 0.85, 0.95, 0.99$), that there is a significant negative effect of patients who were residing around the urban area on their CD4 cell count (see Table 3 and Supplementary information). Table 3 also shows that the negative effect of baseline viral load on the CD4 cell count is higher at the lower quantiles than at the median and higher quantiles (see also, Supplementary information). In addition, R package *aqmm()* sample outputs using CAPRISA 002 AI study data at $\tau = 0.25, 0.75, 0.85$, and 0.99 can be found in Supplementary information.

The variance of the first smooth term ($\hat{\phi}_{Time}$) indicates a stronger penalty on the spline coefficients at $\tau = 0.25, 0.5, 0.75, 0.85$ quantiles than at the bottom and at the top quantiles (Table 4). Similarly, the variance of the second smoother ($\hat{\phi}_{BaselineBMI}$) shows a strong penalty on the spline coefficients at $\tau = 0.25, 0.5, 0.75, 0.85$ quantiles than at the bottom and at more extreme quantiles. Table 4 shows that the random effects' variances have roughly constant variability of subject linear trends across the fitted quantiles (see, also, Supplementary information).

Based on the seven fitted quantile levels ($\tau = 0.05, 0.25, 0.5, 0.75, 0.85, 0.95, 0.99$), Fig. 3 depicts the two estimated smoothed covariate effects on patients' CD4 counts. Patients enrolled in the CAPRISA 002 AI study exhibit nonlinear time effects on CD4 counts that are prominent at all quantile levels. As the quantile increases, its effect becomes stronger. However, it is after several treatment visits that such progress towards higher CD4 counts occurs. Consequently, the progression is slow until about 50 months, then it increases steadily thereafter across all quantile levels (Fig. 3).

Furthermore, overall fit quantile levels, the significant smoothed baseline BMI effect on patients' CD4 counts is roughly constant for patients with a baseline BMI of about 40 but gradually improves from there. Because of this, patients with low BMI need to be monitored carefully before and after HAART initiation. Despite this, physicians should not ignore patients with high BMI. According to our studies and other findings, a plausible explanation may be that BMI may affect drug metabolism and, thus, the progress of HAART and its immunological responses[20,67,68]. Moreover, higher levels of BMI have a more significant effect than lower levels (Fig. 3).

Discussion and conclusion

As a cutting-edge statistical method for modeling percentiles of response variables conditioned on respective covariates, quantile regression is the most widely used. While regression for medians may be seen as more robust than regression for the mean, QR, a generalization of median regression, allows better exploration of data by allowing the modeling of conditional quantiles at low or high extents, such as the 5th and 95th percentiles. As

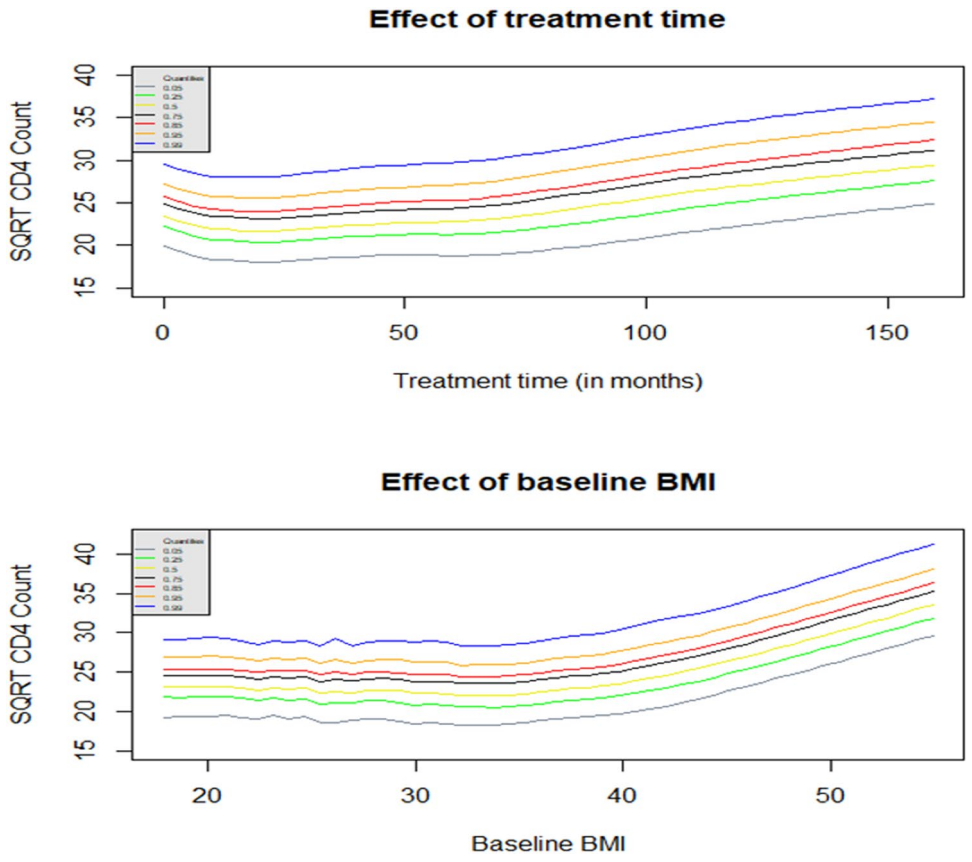


Figure 3. Predicted smoothed covariate effects on the square root CD4 count of HIV-infected patients recurred in the CAPRISA 002 AI study at various quantiles using AQMM.

a result, QR is becoming more common in clinical, biomedical, and other health-related research. Mean-based regression is used to formulate mixed-effects models and their estimated effects on the response variable. In some cases, this centrality-based inference method may not be the optimal method for dealing with the data since the data may not adequately represent their distribution. It has recently been demonstrated that QR has the potential to be extended to a mixed-effects modeling setting, even though QR was initially developed in a univariate setting[48,60,61]. Studies of quantile mixed-effects models have received increasing attention[15,48,60,61,69–76].

Quantile mixed-effects models have been extended to additive models to obtain robust results across various quantile levels of the longitudinal outcome, which brings a rigorous covariates' effect[74–76]. The additive version of the quantile mixed-effects model has gained a great deal of popularity, as discussed above; because it offers an efficient and flexible framework for nonlinear and linear longitudinal forms of data analysis focused on features of the outcome beyond its central tendency[1,4,11,12,47,73,75,76].

In this study, we investigated the effect of multivariate additive quantile mixed models of Geraci[4] on the longitudinal CD4 count of HIV-infected patients across different quantile levels according to parametric and nonparametric covariate effects. By using this recently developed model, robust results are obtained, not only at the central location of the longitudinal outcome that may not be the best place to analyze the data but also at different points of the conditional distribution that gives an inclusive and more complete picture of the parametric as well as the nonparametric covariate effects.

A series of AQMM at $\tau = 0.05, 0.25, 0.5, 0.75, 0.85, 0.95$, and 0.99 were performed, and the results were discussed. According to the results, patients' CD4 count is markedly increased after HAART initiation, and their baseline viral load shows a negative effect on the progression of their CD4 count over time, as we would have expected. All fitted quantiles of the response variable were affected by a significant nonlinear relationship between time and baseline BMI. Study results suggest that, across all fitted quantile levels, the patient's education level does not significantly influence the progression of CD4 counts over time. All but the most extreme quantiles of HIV-positive patients showed a significant difference in the CD4 count regardless of their age. In addition, CD4 cell recovery was found to be significant across all quantiles among patients with a stable sexual partner. Contrary to this, HIV-infected patients with many sexual partners during the treatment period showed a negative effect on CD4 cell count across all fitted quantile levels.

As we expected, the patient's CD4 count increased significantly after HAART was initiated, and their baseline viral load also showed a significant negative effect on the patient's CD4 count over time. Baseline BMI and time were also significant nonlinear effects in our analysis. Further, patients with higher BMIs at baseline have improved CD4 cell count over time after treatment. Despite this, higher BMI patients should not be ignored

clinically. This study instead suggests that BMI can influence drug metabolism and, consequently, the immunological responses to HAART. According to the nonlinear time effect, patients' CD4 counts are not increasing rapidly over time. The growth starts after multiple treatment visits. Hence, the study suggests that HIV patients who are not clinically and immunologically stable on HAART could experience increased risks if exposed to COVID-19, especially if they are not on HAART immediately after HIV exposure.

One can estimate the covariate effects over the grid $\tau \in (0, 1)$ as per the analysis aspects. An investigator, however, should be cautious when using AQMM since the method needs some adjustment to control the estimation algorithm and demands more computing time to estimate the random effects[4]. For instance, for this study, it took 2–3 h to fit the proposed model (13) at a single τ as like Geraci[4]. To overcome this computational burden, Geraci[4] suggested the necessity of further improvement to the AQMM. As the studied data set is an ongoing study, there is a plan to extend AQMM application to genetics in future work since it produces satisfactory results.

Data availability

The dataset used for this study can be obtained by requesting Dr. Nonhlanhla Yende-Zuma (Head of Biostatistics Unit, CAPRISA, Email: Nonhlanhla.Yende@caprisa.org) on reasonable request.

Received: 19 May 2021; Accepted: 19 August 2021
Published online: 09 September 2021

References

1. Fenske, N., Fahrmeir, L., Hothorn, T., Rzehak, P. & Höhle, M. Boosting structured additive quantile regression for longitudinal childhood obesity data. *The International Journal of Biostatistics* **9**(1), 1–18 (2013).
2. Koenker, R. *Quantile Regression* (Cambridge University Press, Cambridge, 2005).
3. Koenker, R. & Bassett, Jr G. Regression quantiles. *Econ. J. Econ. Soc.* **46**(1), 33–50 (1978).
4. Geraci, M. Additive quantile regression for clustered data with an application to children's physical activity. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **68**(4), 1071–1089 (2019).
5. Buchinsky, M. Recent advances in quantile regression models: A practical guideline for empirical research. *J. Hum. Resour* . **33**(1), 88–126 (1998).
6. Koenker, R. & Geling, O. Reappraising medfly longevity: A quantile regression survival analysis. *J. Am. Stat. Assoc.* **96**(454), 458–468 (2001).
7. Peterson, M. D. & Krishnan, C. Growth charts for muscular strength capacity with quantile regression. *Am. J. Prev. Med.* **49**(6), 935–938 (2015).
8. Sherwood, B., Wang, L. & Zhou, X. H. Weighted quantile regression for analyzing health care cost data with missing covariates. *Stat. Med.* **32**(28), 4967–4979 (2013).
9. Yu, K., Lu, Z. & Stander, J. Quantile regression: Applications and current research areas. *J. R. Stat. Soc. Ser. D (The Statistician)* **52**(3), 331–350 (2003).
10. Cade, B. S. & Noon, B. R. A gentle introduction to quantile regression for ecologists. *Front. Ecol. Environ.* **1**(8), 412–420 (2003).
11. Koenker, R. Additive models for quantile regression: Model selection and confidence bandaids. *Braz. J. Probab. Stat.* **25**(3), 239–262 (2011).
12. Fenske, N., Kneib, T. & Hothorn, T. Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. *J. Am. Stat. Assoc.* **106**(494), 494–510 (2011).
13. Yirga, A. A., Ayele, D. G. & Melesse, S. F. Application of quantile regression: Modeling body mass index in Ethiopia. *Open Public Health J.* **11**(1), 221–233 (2018).
14. Winkelmann, R. Reforming health care: Evidence from quantile regressions for counts. *J. Health Econ.* **25**(1), 131–145 (2006).
15. Huang, Q., Zhang, H., Chen, J. & He, M. Quantile regression models and their applications: A review. *J. Biom. Biostat.* **8**, 354. https://doi.org/10.4172/2155-6180.1000354 (2017).
16. Gilchrist, W. *Statistical Modelling with Quantile Functions* (Chapman and Hall/CRC, London, 2000).
17. Wood, S. N. *Generalized Additive Models: An Introduction with R* (CRC Press, London, 2017).
18. Yirga, A. A., Melesse, S. F., Mwambi, H. G. & Ayele, D. G. Negative binomial mixed models for analyzing longitudinal CD4 count data. *Sci. Rep.* **10**(1), 1–15 (2020).
19. AIDSMAP. CD4 cell counts|aidsmap (2021). https://www.aidsmap.com/about-hiv/cd4-cell-counts.
20. Yirga, A. A., Melesse, S. F., Mwambi, H. G. & Ayele, D. G. Modelling CD4 counts before and after HAART for HIV infected patients in KwaZulu-Natal South Africa. *Afr. Health Sci.* **20**(4), 1546–1561 (2020).
21. World Health Organization. *Women, Ageing, and Health: A Framework for Action: Focus on Gender* (2007).
22. World Health Organization. *AIDS Epidemic Update: December 2009* (WHO Regional Office Europe, 2010).
23. UN Women. *Message from UN Women's Executive Director for World AIDS Day, the 1st of December 2014* (2014). https://www.unwomen.org/en/news/stories/2014/12/world-aids-day-2014.
24. AMFAR. *The Foundation for AIDS Research. Statistics: Women and HIV/AIDS* (2015). https://www.amfar.org/about-hiv-and-aids/facts-and-stats/statistics--women-and-hiv-aids/.
25. Whelan, D. Gender and HIV/AIDS: Taking stock of research and programmes. In *UNAIDS* (1999).
26. Kassutto, S. & Rosenberg, E. S. Primary HIV type 1 infection. *Clin. Infect. Dis.* **38**(10), 1447–1453. https://doi.org/10.1086/420745 (2004).
27. Cohen, M. S., Shaw, G. M., McMichael, A. J. & Haynes, B. F. Acute HIV-1 infection. *N. Engl. J. Med.* **364**(20), 1943–1954. https://doi.org/10.1056/NEJMra1011874 (2011).
28. Rosenberg, E. S. *et al.* Immune control of HIV-1 after early treatment of acute infection. *Nature* **407**(6803), 523 (2000).
29. Van Loggerenberg, F. *et al.* Establishing a cohort at high risk of HIV infection in South Africa: Challenges and experiences of the CAPRISA 002 Acute Infection study. *PLOS ONE* **3**(4), e1954 (2008).
30. Garrett, N. *et al.* Acceptability of early antiretroviral therapy among South African women. *AIDS Behav.* **22**(3), 1018–1024 (2018).
31. Mlisana, K. *et al.* Rapid disease progression in HIV-1 subtype C-infected South African women. *Clin. Infect. Dis.* **59**(9), 1322–1331 (2014).
32. Moosa, Y. *et al.* Case report: Mechanisms of HIV elite control in two African women. *BMC Infect. Dis.* **18**(1), 1–7 (2018).
33. Karim, S. A., Williamson, C. & Garrett, N. Viral set point and clinical disease progression: The role of immunological, genetic and viral factors over the course of disease and during antiretroviral therapy. CAP002: Acute Infection Study. (An ongoing study) (2017). Accessed 14 Mar 2021. https://www.caprisa.org/Pages/CAPRISAStudies.
34. Wu, H. & Zhang, J.-T. *Non-parametric Regression Methods for Longitudinal Data Analysis: Mixed-Effects Modeling Approaches* Vol. 515 (Wiley, New York, 2006).
35. Fitzmaurice, G., Davidian, M., Verbeke, G. & Molenberghs, G. *Longitudinal Data Analysis* (CRC Press, London, 2008).

36. Lindsey, J. K. Nonlinear models in medical statistics. *Oxford University Press on Demand* (2001).
37. Davidian, M. & Giltinan, D. M. Nonlinear models for repeated measurement data: An overview and update. *J. Agric. Biol. Environ. Stat.* **8**(4), 387–419 (2003).
38. Fox, J. *Non-parametric Simple Regression: Smoothing Scatterplots, No. 130* (Sage, Thousand Oaks, 2000).
39. Davino, C., Furno, M. & Vistocco, D. *Quantile Regression: Theory and Applications* Vol. 988 (Wiley, New York, 2013).
40. Chaudhuri, P. Global non-parametric estimation of conditional quantile functions and their derivatives. *J. Multivar. Anal.* **39**(2), 246–269 (1991).
41. Hastie, T. J. & Tibshirani, R. J. *Generalized Additive Models* Vol. 43 (CRC Press, London, 1990).
42. Hendricks, W. & Koenker, R. Hierarchical spline models for conditional quantiles and the demand for electricity. *J. Am. Stat. Assoc.* **87**(417), 58–68 (1992).
43. Craig, S. G. & Ng, P. T. Using quantile smoothing splines to identify employment subcenters in a multicentric urban area. *J. Urban Econ.* **49**(1), 100–120 (2001).
44. Koenker, R., Portnoy, S. & Ng, P.. Non-parametric estimation of conditional quantile functions. In Dodge, Y. (Ed) (1992).
45. Koenker, R. Censored quantile regression redux. *J. Stat. Softw.* **27**(1), 1–25 (2008).
46. Cleveland, W. S. & Loader, C. Smoothing by local regression: Principles and methods. In *Statistical Theory and Computational Aspects of Smoothing*. 10–49 (Physica-Verlag HD, 1996).
47. Koenker, R., Ng, P. & Portnoy, S. Quantile smoothing splines. *Biometrika* **81**(4), 673–680 (1994).
48. Liu, Y. & Bottai, M. Mixed-effects models for conditional quantiles with longitudinal data. *Int. J. Biostat.* **5**(1), 28 (2009).
49. Lachos, V. H., Chen, M.-H., Abanto-Valle, C. A. & Azevedo, C. L. Quantile regression for censored mixed-effects models with applications to HIV studies. *Stat. Interface* **8**(2), 203 (2015).
50. Koenker, R. & Hallock, K. F. Quantile regression. *J. Econ. Perspect.* **15**(4), 143–156 (2001).
51. Lin, C. Y., Bondell, H., Zhang, H. H. & Zou, H. Variable selection for non-parametric quantile regression via smoothing spline analysis of variance. *Statistics* **2**(1), 255–268 (2013).
52. He, X., Ng, P. & Portnoy, S. Bivariate quantile smoothing splines. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **60**(3), 537–550 (1998).
53. Li, Y., Liu, Y. & Zhu, J. Quantile regression in reproducing kernel Hilbert spaces. *J. Am. Stat. Assoc.* **102**(477), 255–268 (2007).
54. Rigby, R. A. & Stasinopoulos, D. M. Generalized additive models for location, scale and shape. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **54**(3), 507–554 (2005).
55. Stone, C. J. Additive regression and other non-parametric models. *Ann. Stat.* **13**(2), 689–705 (1985).
56. Breiman, L. & Friedman, J. H. Estimating optimal transformations for multiple regression and correlation. *J. Am. Stat. Assoc.* **80**(391), 580–598 (1985).
57. Der, G. & Everitt, B. S. *Applied Medical Statistics Using SAS* (CRC Press, London, 2012).
58. Xiang, D. Fitting generalized additive models with the GAM procedure. In *SUGI Proceedings* 256–326 (Cary, NC: SAS Institute, Inc., 2001)
59. Ruppert, D., Wand, M. P. & Carroll, R. J. *Semiparametric Regression, No. 12* (Cambridge University Press, Cambridge, 2003).
60. Geraci, M. & Bottai, M. Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics* **8**(1), 140–154 (2007).
61. Galarza Morales, C. E. Quantile regression for mixed-effects models (2015). https://bit.ly/3i7BPyQ.
62. Koenker, R. & Machado, J. A. Goodness of fit and related inference processes for quantile regression. *J. Am. Stat. Assoc.* **94**(448), 1296–1310 (1999).
63. Yu, K. & Moyeed, R. A. Bayesian quantile regression. *Stat. Probab. Lett.* **54**(4), 437–447 (2001).
64. Yu, K. & Zhang, J. A three-parameter asymmetric Laplace distribution and its extension. *Commun. Stat. Theory Methods* **34**(9–10), 1867–1879 (2005).
65. Zuur, A. *et al. Mixed Effects Models and Extensions in Ecology with R* (Springer, New York, 2009).
66. Wood, S. N. Thin plate regression splines. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **65**(1), 95–114 (2003).
67. Palermo, B., Bosch, R. J., Bennett, K. & Jacobson, J. M. Body mass index and CD4+ T-lymphocyte recovery in HIV-infected men with viral suppression on antiretroviral therapy. *HIV Clin. Trials* **12**(4), 222–227 (2011).
68. Li, X. *et al.* Predictive effects of body mass index on immune reconstitution among HIV-infected HAART users in China. *BMC Infect. Dis.* **19**(1), 1–9 (2019).
69. Galvao, A. F. Jr. Quantile regression for dynamic panel data with fixed effects. *J. Econ.* **164**(1), 142–157 (2011).
70. Fu, L. & Wang, Y.-G. Quantile regression for longitudinal data with a working correlation model. *Comput. Stat. Data Anal.* **56**(8), 2526–2538 (2012).
71. Lipsitz, S. R., Fitzmaurice, G. M., Molenberghs, G. & Zhao, L. P. Quantile regression methods for longitudinal data with drop-outs: Application to CD4 cell counts of patients infected with the human immunodeficiency virus. *J. Roy. Stat. Soc. Ser. C (Appl. Stat.)* **46**(4), 463–476 (1997).
72. Geraci, M. & Bottai, M. Linear quantile mixed models. *Stat. Comput.* **24**(3), 461–479 (2014).
73. Galarza, C. E., Lachos, V. H. & Bandyopadhyay, D. Quantile regression in linear mixed models: A stochastic approximation EM approach. *Stat Interface* **10**(3), 471 (2017).
74. Yue, Y. R. & Rue, H. Bayesian inference for additive mixed quantile regression models. *Comput. Stat. Data Anal.* **55**(1), 84–96 (2011).
75. Sriram, K., Shi, P. & Ghosh, P. A Bayesian semiparametric quantile regression model for longitudinal data with application to insurance company costs. In *IIM Bangalore Research Paper* 355 (2011).
76. Huang, Y. Quantile regression-based Bayesian semiparametric mixed-effects models for longitudinal data with non-normal, missing and mismeasured covariate. *J. Stat. Comput. Simul.* **86**(6), 1183–1202 (2016).

Acknowledgements

We gratefully acknowledge CAPRISA for giving us access to the CAPRISA 002: Acute Infection Study data. CAPRISA is funded by the National Institute of Allergy and Infectious Diseases (NIAID), National Institutes for Health (NIH), and U.S. Department of Health and Human Services (Grant: AI51794). The authors would also like to thank Dr. Nonhlanhla Yende-Zuma (Head of Biostatistics unit at CAPRISA) for her cooperation, assistance, and technical support.

Author contributions

A.A.Y. obtained the data, did the analysis, and prepared the manuscript. A.A.Y., S.F.M., H.G.M., and D.G.A. planned the research problem. All authors deliberated on the results and consequences and commented on the paper at all stages. All authors contributed extensively to the work presented in this manuscript. All authors read and ratified the ultimate manuscript.

Funding
This work was supported through the DELTAS Africa Initiative and the University of KwaZulu-Natal. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [Grant 107754/Z/15/Z], DELTAS Africa Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB) programme and the UK government.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-97114-9.

Correspondence and requests for materials should be addressed to A.A.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

BMC Infectious Diseases

RESEARCH ARTICLE

Open Access

Application of quantile mixed-effects model in modeling CD4 count from HIV-infected patients in KwaZulu-Natal South Africa

Ashenafi A. Yirga[1*], Sileshi F. Melesse[1], Henry G. Mwambi[1] and Dawit G. Ayele[2]

Abstract

Background: The CD4 cell count signifies the health of an individual's immune system. The use of data-driven models enables clinicians to accurately interpret potential information, examine the progression of CD4 count, and deal with patient heterogeneity due to patient-specific effects. Quantile-based regression models can be used to illustrate the entire conditional distribution of an outcome and identify various covariates effects at the respective location.

Methods: This study uses the quantile mixed-effects model that assumes an asymmetric Laplace distribution for the error term. The model also incorporated multiple random effects to consider the correlation among observations. The exact maximum likelihood estimation was implemented using the Stochastic Approximation of the Expectation–Maximization algorithm to estimate the parameters. This study used the Centre of the AIDS Programme of Research in South Africa (CAPRISA) 002 Acute Infection Study data. In this study, the response variable is the longitudinal CD4 count from HIV-infected patients who were initiated on Highly Active Antiretroviral Therapy (HAART), and the explanatory variables are relevant baseline characteristics of the patients.

Results: The analysis obtained robust parameters estimates at various locations of the conditional distribution. For instance, our result showed that baseline BMI (at $\tau = 0.05$: $\widehat{\beta}_4 = 0.056$, p−value < 0.0064; at $\tau = 0.5 : \widehat{\beta}_4 = 0.082$, p−value < 0.0025; at $\tau = 0.95 : \widehat{\beta}_4 = 0.145$, p−value < 0.0000), baseline viral load (at $\tau = 0.05$: $\widehat{\beta}_5 = -0.564$, p−value < 0.0000; at $\tau = 0.5 : \widehat{\beta}_5 = -0.641$, p−value < 0.0000; at $\tau = 0.95 : \widehat{\beta}_5 = -0.739$, p−value < 0.0000), and post-HAART initiation (at $\tau = 0.05$: $\widehat{\beta}_6 = 1.683$, p−value < 0.0000; at $\tau = 0.5 : \widehat{\beta}_6 = 2.560$, p−value < 0.0000; at $\tau = 0.95 : \widehat{\beta}_6 = 2.287$, p−value < 0.0000) were major significant factors of CD4 count across fitted quantiles.

Conclusions: CD4 cell recovery in response to post-HAART initiation across all fitted quantile levels was observed. Compared to HIV-infected patients with low viral load levels at baseline, HIV-infected patients enrolled in the treatment with a high viral load level at baseline showed a significant negative effect on CD4 cell counts at upper quantiles. HIV-infected patients registered with high BMI at baseline had improved CD4 cell count after treatment, but physicians should not ignore this group of patients clinically. It is also crucial for physicians to closely monitor patients with a low BMI before and after starting HAART.

Keywords: Quantile regression, Quantile mixed model, Stochastic approximation of the expectation maximization, Asymmetric Laplace distribution, CD4 count, CAPRISA

*Correspondence: ashu3argaw@gmail.com; 216065934@stu.ukzn.ac.za
[1] School of Mathematics, Statistics, and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, Private Bag X01, Scottsville 3209, South Africa
Full list of author information is available at the end of the article

Background

CD4 cell counts indicate a sign of the wellbeing of the immune system for an individual. It also provides information about disease progression. The number of CD4

cells of an individual who does not have HIV could be somewhere in the range of 500 to 1500 cells/mm^3. "Individuals living with HIV who have a CD4 count above 500 cells/mm^3 are usually in good health. Individuals living with HIV who have a CD4 cell count less than 200 cells/mm^3 are at high risk of developing severe sickness" [1]. HIV therapy is recommended for all individuals infected with HIV. It is particularly critical for patients with low CD4 count to preferably starting treatment sooner rather than later, under the current WHO recommendation for individuals who test HIV positive [2].

The classical regression model about the mean has been the commonly applied statistical procedure to depict the effects of explanatory variables for continuous outcomes. Despite this, such results based on a fixed location of the response distribution may not be relevant in many areas, and sometimes the fields of application are diverse. Numerous investigators, economic experts, monetary stakeholders, clinicians, and legislators have revealed a growing interest in group differences across the whole population instead of relying solely on the average [3–6]. Another approach to studying the central location is median regression. The median regression approach is robust to the manifestation of outliers and when the error distribution is not correctly specified [3, 7].

Quantile regression (QR) was popularized by Koenker and Bassett [7]. It is an extension of median regression to examine the covariates' influence on different quantiles of the entire response distribution. Fixed effects could have different impacts across various quantile levels. QR allows for a wide range of applications, for example, investigating the 5th or 25th percentile (lower quantiles) of the response (e.g., CD4 count distribution of HIV infected patient) might be of interest in studying patients with lower CD4 cell counts, where individuals are at higher risk of developing illnesses. Therefore, it is important to study the response distribution across all quantiles (e.g., at different CD4 count distribution), rather than only the central tendency, such as in mean or median regression.

In recent years, mixed quantile regression models have become a widely used technique in statistical studies. By using quantile-based regression model, it is possible to examine the location, scale, and shape of the distribution of responses to get an idea of how the covariates affect the distribution of responses. It is also more robust to outliers when compared to conventional mean regression and is invariant to monotonic transformations. There is no need to make any Gaussian assumptions concerning the response with quantile regression, and further it is capable of handling heavy-tailed and asymmetric data. As a result, CD4 count can be modeled very well using this method.

Many longitudinal studies gather a great deal of information about repeated measures that are crucial for analyzing disease progression in clinical studies. For example, repeated counts of CD4 cells are vital to HIV/AIDS monitoring; for instance, low levels of CD4 counts are signs of serious viral load accumulation, disease progression, and the need for therapy intervention. Physicians also use them to identify the advantages of medical involvement and the risk factors that may lead to poor outcomes. In practical statistics, mixed-effects models have become quite popular. As a result of their ability to handle both between-subjects and within-subjects variability in longitudinal data, they are often used to analyze longitudinal data [8]. Mixed-effects models and their estimated effects are formulated on the response variable via mean regression, accounting for between-subject heterogeneity through normally distributed subject-specific random effects and random errors. Mixed-effects models have been studied extensively (see, for example, [8–12]). There are also various strategies applicable to handle longitudinal data, for instance, generalized estimating equations which are conceptually generalized linear mixed-effects models. However, all these techniques limit the investigation of variations between subjects based on the mean of the response variable, and the latter utilize parametric models based on the normal distributional assumption [3].

Moreover, in some cases, it could be challenging to obtain appropriate transformation to normality for the response variable, or some strategy to account for outliers may be required. An adequate solution to all these issues is given by concentrating on the conditional quantiles of the longitudinal outcome [13]. "Conditional QR methods, dealing with the complete conditional distribution of the response variable, have been developed to grant an analysis of variable effects at any subjective quantiles of the response distribution. Furthermore, QR techniques do not require any distributional assumption on the error; besides that, the error term has a zero-conditional quantile, like the ALD" [14].

The QR method was initially developed in a univariate setting. However, the large amount of longitudinal data has recently dictated its extension into a mixed-effects modeling system by either the distribution-free way [15–17] or the likelihood-based way in most cases following the ALD [18–21]. The likelihood-based quantile mixed model additionally makes use of different parametric distributions, such as an infinite mixture of Gaussian densities [22] and a direct parametric maximum likelihood (ML) approach [23]. The distribution-free approaches that consist of fixed-effects and weighted generalized

estimating equations consider the use of *independent* estimating equations that ignore correlations between repeated measurements which leads to loss of efficiency [5, 17, 20]. Meanwhile, Geraci and Bottai [19] suggested a likelihood-based QR model for longitudinal data that accounts for within-subject dependence by incorporating subject-level random effects and modeling the residual distribution with an ALD. Liu and Bottai [24] developed a likelihood-based method to estimate parameters of conditional quantile functions with random effects by incorporating an ALD for the random error term that is not restricted to be normal. The within-subject correlation is taken into consideration by incorporating random effects to get unbiased parameter estimates [24]. The application of QR for mixed-effects models has received increasing consideration in wide-ranging areas of study, including marine biology, environmental science, cardiovascular disease, and ophthalmology [19, 20, 25–28]. Following the version of the quantile mixed model of Galarza [18], this study aims to model the longitudinal CD4 count of HIV-infected patients using quantile mixed-effects models using the likelihood-based function that uses ALD for the error term. The study employs data from the CAPRISA 002 AI Study. In this study, we will demonstrate how quantile mixed model can be used to estimate covariate effects at different locations of the conditional distribution that communicates a wide-range and more complete picture of the effects.

Methods

Data description

This study used the Centre of the AIDS Programme of Research in South Africa (CAPRISA) 002 Acute Infection (AI) Study data conducted at the Doris Duke Medical Research Institute (DDMRI) at the Nelson R Mandela School of Medicine of the University of KwaZulu-Natal (UKZN) in Durban, South Africa [29–33]. CAPRISA started the CAPRISA 002 AI study between August 2004 and May 2005 by enrolling women who are at high risk of HIV infection for follow-up with an intense on-going examination to help estimate HIV infection rates within the study, including providing intense aftercare advice to those dropping out prematurely, the careful follow-up to study disease progression, and CD4 count and viral load evolution [29–33]. Detail description of the design, development, and procedures of the CAPRISA 002 AI study population can be found here [29, 30].

When an infected person's body indicates symptoms of being incapable of adequately controlling the virus and their CD4 count drops below a specific cut point, they were initiated on therapy. A deficient level of CD4 count causes the weak immune system of an HIV-infected person. In the absence of treatment or without viral suppression, the person is susceptible to opportunistic infections (OIs). This increases the risk of the new and ongoing Coronavirus Disease 2019 (COVID-19) infections and underlying illnesses [31–33]. HAART is an effective way of preventing these infections and diseases. By suppressing and preventing the virus from making copies of itself, HAART aims to decelerate or prevent the progression to AIDS and loss of life for HIV-infected people. The body's immune system is less damaged, and HIV infection complications are decreased when the level of the virus in the blood is low or "undetectable" through HAART [31–33]. This is also significantly reducing the likelihood of transmitting HIV to partners.

The HIV/AIDS epidemic and other sexually transmitted diseases severely impact human health, especially the well-being of women and young girls [31–33]. "The consequences of HIV/AIDS stretch beyond women's health to their part as moms and caregivers and their commitment to their families' economic support. The social, development, and health consequences of HIV/AIDS and other sexually transmitted illnesses ought to be considered from a gender perspective" [34–36]. Apart from sex-specific issues, HIV therapy algorithms for women are similar to that of men [31]. The interaction between the clinician and the changing HIV epidemiology will provide the clinician with a technique to identify patients at high risk of HIV infection and clarify which rules should be applied to avoid sequential HIV transmission [31–33]. Although ART suggestions are the same for all patients, the study of CD4 count of HIV-infected patients, in conjunction with individual differences, will help clinicians to get through and interpret potential information precisely due to patient specific-specific effects [31, 33, 37–39].

Quantile mixed-effects model

Quantile regression (QR) is an advanced statistical technique to study the predictors' heterogeneous effects at the conditional distribution of the outcome. Instead of modeling only the mean value like the conventional regression methods, quantile regression enables more fully to explore the data by modeling the conditional quantiles, for example, the 5th and 95th percentiles of the response distribution [33]. For these reasons, it has become more prevalent in several epidemiological and economics studies. For instance, Yirga et al. [40] studied how children's BMI varies with age and other factors using quantile regression. There are several other applications of quantile regression based on uncorrelated data, among which public health, bioinformatics, health care, environmental science, ecology, microarray data analysis, and survival data analysis [13, 41–51].

The quantile level is frequently signified by the Greek letter τ, and the conditional quantile of y given x is often written as $Q_\tau(y|x)$. The quantile level τ is the probability $\Pr[y \le Q_\tau(y|x)]$, and it is the value of y below which the proportion of the conditional response population is τ. For a random variable y with a probability distribution function $F(y) = Pr(Y \le y)$, the τ quantile of y is defined as the inverse function $Q(\tau) = inf\{y : F(y) \ge \tau\}$, $\tau\epsilon(0,1)$. Particularly, the median is $Q(0.5)$. Let y_i denote a scalar response variable with conditional cumulative distribution function F_{y_i}, whose shape is unspecified and $\boldsymbol{x}_i$ the corresponding covariates vector of dimension $k \times 1$ for subject $i, i = 1, \ldots, n$. Then, following Koenker and Basset (1978), the τth$(0 < \tau < 1)$ quantile regression modeled is written as $Q_\tau(y_i|\boldsymbol{x_i}) = \boldsymbol{x_i'\beta_\tau}$, where $Q_\tau(y_i|\boldsymbol{x_i}) \equiv F_{y_i}^{-1}(\bullet)$, which is the quantile function (or the inverse cumulative distribution function) of y_i given $\boldsymbol{x_i}$ estimated at τ, and $\boldsymbol{\beta}_\tau$ is a column vector of regression parameters corresponding to the τth quantile. On the other hand, this expression can be written as

$$Q_\tau(y|\boldsymbol{x_i}) = \boldsymbol{x_i'\beta_\tau} + \varepsilon_i, \text{with } Q_{\varepsilon_i}(\tau|\boldsymbol{x}_i) = 0, \tag{1}$$

where ε_i is the error term whose distribution (with density $f_\tau(\bullet)$) is restricted to have the τth quantile to be zero, that is, $\int_{-\infty}^{0} f_\tau(\varepsilon_i)d\varepsilon_i = \tau$ [24, 52]. "The error density $f_\tau(\bullet)$ is often left unspecified in the classical literature" [52]. Thus, the estimator $\widehat{\boldsymbol{\beta}}_\tau$ proceeds through *linear programming (LP)* by minimizing

$$\widehat{\boldsymbol{\beta}}_\tau = \underset{\beta \epsilon R^p}{\operatorname{argmin}} \sum\nolimits_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x_i'\beta_\tau}), \tag{2}$$

where $\rho_\tau(\bullet)$ is the so called loss (or check) function defined by $\rho_\tau(u) = u(\tau - I\{u < 0\})$ with u being a real number and $I\{\bullet\}$ is the indicator function. Thus, $\widehat{\boldsymbol{\beta}}_\tau$ is called the τth quantile regression estimate [5, 13, 43, 53]. The parameter $\boldsymbol{\beta}_\tau$ and its estimator $\widehat{\boldsymbol{\beta}}_\tau$ depends on the quantile τ, because of different choices of τ estimate different values of β [24]. For this reason, the interpretation of $\boldsymbol{\beta}_\tau$ is specific to the quantile being estimated, the intercept term denotes the baseline predicted value of the response at specific quantile τ, while each coefficient can be interpreted as the rate of change of the τth response quantile per unit change in the value of the corresponding predictor variable (ith regressor) keeping all the other covariates constant.

The objective function of the conditional quantile estimator, $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\tau}}$, in Eq. (2) proceeds by minimizing

$$\begin{aligned} H(\boldsymbol{\beta_\tau}) &= \sum_i \tau|\varepsilon_i| + \sum_i (1-\tau)|\varepsilon_i| \\ &= \sum_{i:y_i \ge \boldsymbol{x_i'\beta_\tau}}^{n} \tau|y_i - \boldsymbol{x_i'\beta_\tau}| \\ &\quad + \sum_{i:y_i < \boldsymbol{x_i'\beta_\tau}}^{n} (1-\tau)|y_i - \boldsymbol{x_i'\beta_\tau}|, 0 < \tau < 1 \end{aligned} \tag{3}$$

where $i : y_i \ge \boldsymbol{x_i'\beta}$ for under prediction, and $i : y_i < \boldsymbol{x_i'\beta}$ for overprediction [5]. Since the above objective function is nondifferentiable, the gradient optimization methods are not applicable; instead, *LP* methods can be used to obtain $H(\boldsymbol{\beta_\tau})$ [41, 54, 55]. For more details and a summary of quantile regression, see, for example, Davino et al. [3], Konker and Basset [7], Konker [13], Buchinsky [41], Koenker and Hallock [43], or Yu et al. [49].

As the check function $(\rho_\tau(\bullet))$ in Eq. (2) is not differentiable at zero, we cannot extract specific solutions to the minimization problem. Hence, *LP* procedures are often used to achieve a relatively fast computation of $H(\boldsymbol{\beta_\tau})$ [52, 56]. A natural link between minimization of the quantile check function and ML theory is given by the assumption that the error term in Eq. (1) follows an ALD [53, 57]. A connection between the minimization of the sum in Eq. (2) and the ML theory is provided by ALD [58]. Other forms of Laplace distribution were summarized by Kotz et al. [59] and Kozubowski and Nadarajah [60]. ALD that is closely associated with the loss function for QR has been examined in several works of literature [19, 24, 52, 57, 58].

The conventional QR is based on the median, or other quantile levels, by assuming a continuous or Gaussian distribution. QR has been extended to count regression, which is a special case of the discrete variable model [55, 56, 61–64]. However, the distribution function of a discrete random variable is not continuous, and the objective function of the conditional quantile $Q_\tau(y|\boldsymbol{x})$ for a discrete distribution cannot be a continuous function of $\boldsymbol{x}$ such as $\exp(\boldsymbol{x'\beta})$ [61]. Machado and Silva [64] overcome this restriction by developing a continuous random variable whose quantiles have a one-to-one relation with the quantiles of y, a count variable. When count data consists of severe outliers or multiple distributional components that do not reflect a known underlying probability distribution, quantile count models may be a useful alternative. Furthermore, QR models all of the quantiles of the discrete distribution and covers the entire range of counts [62]. Detailed discussions about quantile count models for independent data are available in Winkelmann [61], Machado and Silva [64], Hilbe [62, 63], Cameron and Trivedi [55, 56], and a recent application of this model can be found in Winkelmann [65] and Miranda [66].

Mixed-effects models characterize an ordinary and conventional type of regression methods used to examine data coming from longitudinal studies. The general linear mixed-effects model is defined as

$$\boldsymbol{Y}_i = \boldsymbol{X}_i^{'}\boldsymbol{\beta} + \boldsymbol{Z}_i^{'}\boldsymbol{u}_i + \varepsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, n_i,$$

where $\boldsymbol{Y}_i$ is the $n_i \times 1$ vector of the response variable, $\boldsymbol{X}_i^{'}$ is a known $n_i \times p$ design matrix that includes covariates for the fixed effects, $\boldsymbol{\beta}$ is $p \times 1$ vector of population-averaged fixed-effects, $\boldsymbol{Z}_i^{'}$ with the dimension of $n_i \times r$ known design matrix for random effects, $\boldsymbol{u}_i$ is $r \times 1$ vector of random effects, $\boldsymbol{u}_i \sim N(0, \boldsymbol{\Sigma}_u)$, and ε_{ij} is the independent and identically distributed random errors, $\varepsilon_{ij} \sim N(0, \sigma^2)$. Thus, the τth quantile linear mixed-effects model, which were developed by Geraci and Bottai [20] as an extension of the QR model with a random intercept of Geraci and Bottai [19], of a continuous response $\boldsymbol{Y}_i$, has the form

$$Q_\tau\left(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{u}_i\right) = \boldsymbol{x}_{ij}^{'}\boldsymbol{\beta}_\tau + \boldsymbol{z}_{ij}^{'}\boldsymbol{u}_i + \varepsilon_{\tau,ij}, 0 < \tau < 1 \quad (4)$$

where y_{ij} is the response of subject i at jth measurement, $\boldsymbol{x}_{ij}$ indicates covariate vector of ith subject at jth measurement for fixed effects, $\boldsymbol{z}_{ij}$ indicates covariate vector of ith subject at jth measurement for the random effects $\boldsymbol{u}_i$, and random errors $\varepsilon_{\tau,ij} \sim ALD(0, \sigma, \tau)$, which are also dependent on τ. $\boldsymbol{\beta}_\tau$ is the coefficient of fixed-effects corresponding to the τth quantile, and the response variable y_{ij}, conditional on $\boldsymbol{x}_{ij}$, $\boldsymbol{u}_i$, for $i = 1, \ldots, n, j = 1, \ldots, n_i$ and σ are assumed to be conditionally independently distributed as ALD with the density given by

$$f\left(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{u}_i, \sigma\right) = \frac{\tau(1-\tau)}{\sigma} exp\left\{-\rho_\tau\left(\frac{y_{ij} - \boldsymbol{x}_{ij}^{'}\boldsymbol{\beta}_\tau - \boldsymbol{z}_{ij}^{'}\boldsymbol{u}_i}{\sigma}\right)\right\}. \quad (5)$$

The random effects ($\boldsymbol{u}_i$'s) are assumed to be distributed as $\boldsymbol{u}_i \overset{iid}{\sim} N_r(0, \boldsymbol{\Psi})$, where the dispersion matrix $\boldsymbol{\Psi} = \boldsymbol{\Psi}(\boldsymbol{\alpha})$ relies on unknown and reduced parameters $\boldsymbol{\alpha}$, which is the distinct elements of $\boldsymbol{\Psi}$, and the random errors $\varepsilon_{ij} \sim ALD(0, \sigma)$ [18, 52]. Then a likelihood for y_{ij} at τth quantile is

$$L(\boldsymbol{\beta}_\tau, \sigma, \tau) = \frac{\tau^n(1-\tau)^n}{\sigma^n} exp\left\{-\sum_{i=1}^{n}\sum_{j=1}^{n_i}\rho_\tau\left(\frac{y_{ij} - \boldsymbol{x}_{ij}^{'}\boldsymbol{\beta}_\tau - \boldsymbol{z}_{ij}^{'}\boldsymbol{u}_i}{\sigma}\right)\right\} \quad (6)$$

Based on the likelihood of conditional quantile of y_{ij}, it is suggested that the maximization of the likelihood in Eq. (5) with respect to the parameter $\boldsymbol{\beta}_\tau$ is equivalent to the minimization of the loss function in Eq. (7). Thus, we can estimate the coefficient of fixed-effects corresponding to the τth quantile ($\boldsymbol{\beta}_\tau$) by minimizing the objective function of Eq. (6), which can be expressed as

$$H^*(\boldsymbol{\beta}_\tau) = \min_{\boldsymbol{\beta}_\tau}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\rho_\tau\left(\frac{y_{ij} - \boldsymbol{x}_{ij}^{'}\boldsymbol{\beta}_\tau - \boldsymbol{z}_{ij}^{'}\boldsymbol{u}_i}{\sigma}\right) \quad (7)$$

More details regarding the estimation process of quantile mixed-effects models are available here [18, 19, 24, 58].

Stochastic approximation of the expectation maximization

The study examines quantile regression for linear mixed-effects models (QR-LMM) of Galarza [18] that follows the SAEM algorithm for determining exact ML estimates of the fixed-effects and the general variance–covariance matrix $\boldsymbol{\Sigma}_\tau = \boldsymbol{\Sigma}(\boldsymbol{\theta}_\tau)$ of the random effects parameters for the specific quantile. The Expectation–Maximization algorithm, also known as the EM algorithm, which was suggested by Dempster et al. [67], is a popular technique for iterative computation of ML estimates when the observations are regarded as incomplete data, which incorporates the ordinary or standard elements of missing data; however, it is much broader than that [68]. There are two steps in every iteration of the EM algorithm: an expectation, or E-step, followed by a maximization (M-step). "In the former action, the incomplete data are estimated given the observed data and current estimate of the model parameters under the assumption of missing at random (MAR) for the incomplete data. In the later step, the likelihood function is maximized under the assumption that the incomplete/missing data is known" [67]. The detailed explanations of these processes, their related analytical clarifications for successively more common sorts of models, and the basic theory underlying the EM algorithm are given by Dempster et al. [67]. A book devoted entirely to the general formulation of the EM algorithm and its basic properties and applications has been provided by McLachlan and Krishnan [68]. Moreover, the success of the EM algorithm is well documented and can be found in numerous statistical literature.

Even though the EM algorithm is popular, Delyon et al. [69] pointed out that, in some situations, it is not applicable due to the fact that the E-step cannot be carried out in a closed-form. To deal with these issues, Delyon et al. [69] presented a simulation-based SAEM algorithm based on stochastic approximation (SA) as an elective to the MCEM, standing for Monte Carlo EM. "While the MCEM requires a consistent increment of the simulated data and regularly a substantial number of simulations, the SAEM versions guarantee convergence with a fixed and/or small simulation size" [69–71]. The SAEM algorithm restores the E-step of the

EM algorithm by one iteration of a stochastic (probabilistic) approximation procedure, whereas the M-step is consistent [71]. The E- and M-steps of the EM and SAEM procedures are highlighted as follows.

Let $\updownarrow_o(\widehat{\theta}) = \log f(Y_{obs};\theta)$ denotes the maximization of log-likelihood function based on the observed data (Y_{obs}), and given q represents missing data, $Y_{com} = (Y_{obs}, q)'$ denotes the complete data with observed and missing data, thus $\updownarrow_c(Y_{com};\theta)$ be the complete log-likelihood function, and $\widehat{\theta}_k$ indicates the evaluation of θ at the kth iteration. Then the EM algorithm with missing data that maximizes $\updownarrow_c(Y_{com};\theta) = \log f(Y_{obs}, q;\theta)$ iteratively and converges to a stationary point of the observed likelihood under mild regularity conditions [18, 71], go through in two steps:

- E-step: Consists computing of the conditional expectation of $\updownarrow_c(Y_{com};\theta)$.

$$S\left(\theta|\widehat{\theta}_k\right) = E\left\{\updownarrow_c(Y_{com};\theta)|Y_{obs},\widehat{\theta}_k\right\}$$

- M-step: Computes the parameter values $\widehat{\theta}_{k+1}$ by maximizing $S\left(\theta|\widehat{\theta}_k\right)$ with respect to θ.

The SAEM algorithm, on the other hand replaces the E-step by stochastic approximation, presented by Galarza [18] summarized as follows:

- Simulation (E-step): Generate $q(\updownarrow_o, k)$ sample (simulation of the missing data at iteration k), $\updownarrow = 1, 2, \ldots, m$, from the conditional distribution of the missing data $f\left(q|\theta_{k-1}, Y_{obs}\right)$.
- Stochastic approximation: Update $S\left(\theta|\widehat{\theta}_k\right)$ according to

$$S\left(\theta|\widehat{\theta}_k\right) = S\left(\theta|\widehat{\theta}_{k-1}\right) + \delta_k\left[\frac{1}{m}\sum_{\updownarrow=1}^{m}\updownarrow_c\left(Y_{obs}, q\left(\updownarrow_o, k\right)|\widehat{\theta}_k;\theta\right) - S\left(\theta|\widehat{\theta}_{k-1}\right)\right]$$

- M-step: Maximize $\widehat{\theta}_k$ according to

$$\widehat{\theta}_{k+1} = \underset{\theta}{\operatorname{argmax}} S\left(\theta|\widehat{\theta}_k\right),$$

this is equivalent to finding $\widehat{\theta}_{k+1}\epsilon\boldsymbol{\Theta}$ such that $S\left(\widehat{\theta}_{k+1}\right) \geq S\left(\widehat{\theta}_k\right)\forall\theta\epsilon\boldsymbol{\Theta}$, where δ_k is a smoothing parameter (a sequence of decreasing non-negative numbers) as given by Kuhan and Lavielle [72, 73], and m is the number of simulations suggested to be less than or equal to 20 [18]. The choice of δ_k recommended by Galarza [18] is given as follows:

$$\delta_k = \begin{cases} 1 for\ 1 \leq k \leq cW \\ \frac{1}{k-cW} for\ cW+1 \leq k \leq W, \end{cases}$$

where $c\epsilon(0,1)$ is a cut point that regulates the percentage of initial iterations with no memory, and W is the maximum number of iterations.

For more points of interest, however, see Jank [70], Meza et al. [71], or Kuhn and Lavielle [72, 73]. Furthermore, details of these algorithms for estimating the parameters of the QR-LMM are presented by Galarza [18] and Galarza et al. [21]. "The SAEM algorithm has proven to be more effective for computing the ML estimates in mixed-effects models due to the reusing of simulations from one iteration to the next in the smoothing phase of the algorithm" [18, 71–73]. The SAEM algorithm is employed in the R package *qrLMM*.

Results

CD4 cells are the utmost target of HIV infection, and the CD4 count is used as a health marker for an individual's immune system. Hence, it is of interest to investigate the evolution of CD4 count and disease progression of an individual over time, especially for HIV-infected patients. Consequently, this study analyzes the repeated CD4 count of HIV-positive patients registered in the CAPRISA 002 AI study by employing a parametric quantile regression mixed-effects model based on the asymmetric Laplace distribution. The CAPRISA 002 AI study dataset consists of repeated CD4 count measurements and some other covariates of 235 individuals. There were a total of 7019 observations from the 235 women; each subject was measured several times, ranging from 2 to 61 months, with a median equal to 29. Table 1 illustrates a summary of the patients' baseline characteristics. The patients' age at enrollment ranges from 18 to 59, with the median age being 25 years. $Q_{0.05}$, which is a value that has 5% of the observation smaller or equal to it, indicates that 5% of the patients had a square root of CD4 count below or equal to 16.4 at enrollment. $Q_{0.95}$ is similarly a value that shows 95% of the observation smaller or equal to it; said otherwise, 5% of the patients are greater than it. Therefore, Table 1 indicates 5% of the study participant had a square root CD4 count greater than 31.4 at enrollment. Moreover, the study participants had a mean BMI of 28.93 with minimum and maximum BMI of 17.89 and 54.89 at baseline. The median log baseline VL of the patients was 10.26 with minimum and maximum

Table 1 Summary of patients' baseline characteristics

Variable	Analysis						
	Mean	Median	Minimum	Maximum	$Q_{0.05}$	$Q_{0.95}$	IQR
SQRT_CD4 count	23.44	22.89	13.49	39.49	16.40	31.40	5.78
Baseline BMI	28.93	27.24	17.89	54.89	20	43.70	9.66
Log_Baseline VL	10.09	10.26	0 (undetected)	15.52	6.19	13.13	2.91
Age at baseline	27.15	25	18	59	20	41	8

log baseline VL of 0 (Not detected) and 15.52, respectively (IQR = 2.91). Additional features on this dataset can be found here [29, 30, 32, 33]. We analyze this dataset intending to explain the different conditional distribution of the square-root-transformed CD4 count as a function of sets of covariates of interest through modeling a framework of response quantiles.

The linear mixed-effects model form of the data can be specified as:

$$\begin{aligned} y_{ij} =& \beta_1 + \beta_2 t_i + \beta_3 \sqrt{t_i} + \beta_4 BMI_i \\ &+ \beta_5 LVL_i + \beta_6 ART_i + \beta_7 Age_i + b_{1i} \\ &+ b_{2i} t_i + b_{3i} \sqrt{t_i} + \varepsilon_{ij}, \end{aligned}$$

where y_{ij} is the transformed continuous form of CD4 count ($\sqrt{CD4\,count}$) at the jth time point for the ith subject, t is the time measured in months from the start of the study, BMI indicates the patient's baseline BMI, LVL = log of baseline VL, ART is the dichotomous HAART initiation (0 = pre-HAART, 1 = post-HAART), Age is patient's age at baseline, b_{1i} indicates the random intercept, b_{2i} and b_{3i} indicates the random slopes (for time and square root time respectively) for subject i, and ε_{ij} the measurement error term, assuming ALD, for 235 subjects.

The information criteria are used to compare four models. The models were compared based on the 0.5th quantile (median regression). The linear quantile mixed-effects model with random intercept and slopes (Model 4, see Table 2) was selected as the best model because the chosen model achieved the smallest Akaike information criteria (AIC), Bayesian information criteria (BIC), Hannan–Quinn information criteria (HQC), and the largest Log-likelihood (LL) (see Table 2). Therefore, we examine the square-root-transformed CD4 count of HIV-infected patients as a response while accounting for Baseline BMI, age, log baseline VL, and HAART initiation as predictor variables across various quantiles based on Model 4 (Table 3). A series of QR-LMM at $\tau = 0.05, 0.25, 0.5, 0.75, 0.85,$ and 0.95 are performed to get a complete picture of the effects (see, Table 3, and Additional files 1, 2).

Table 2 Comparison of random effects models for QR-LMM at the 0.5th quantile

Random effects	AIC	BIC	HQC	LL
Model 1	39,670.99	39,725.84	39,689.89	− 19,827.50
Model 2	35,072.84	35,141.41	35,096.47	− 17,526.42
Model 3	35,726.22	35,794.79	35,749.85	− 17,853.11
Model 4	33,685.92	33,781.91	33,718.99	− 16,828.96

Random effect models that were examined for the analysis

Model 1: Time (Random slope model)
Model 2: Intercept, Time (Random intercept and slope model)
Model 3: Time, $\sqrt{Time}$ (Random slopes model)
Model 4: Intercept, Time, $\sqrt{Time}$ (Random intercept and slopes model)

As can be observed from Table 3, the intercept (β_1), which is the predicted value of the square-root-transformed CD4 count keeping all the other covariates zero, differ significantly across the quantiles, while time (β_2), square root of time (β_3), baseline BMI (β_4), the log of baseline VL (β_5), and post HAART initiation (β_6) significantly affect the CD4 count across all quantiles. In addition, although age (β_7) is found to have a positive and almost constant influence on the CD4 count across all quantiles, its effect is non-significant (Table 3). We can also see from Table 3; there is a remarkable positive effect of baseline BMI on square root CD4 cell count ($\sqrt{CD4\,count}$) from low quantiles to higher quantiles. Whereas, from low to more upper quantiles, the negative effect of baseline VL on the count of CD4 cells increases gradually. This indicates that when the VL at enrollment is high (baseline VL at higher quantiles), its negative

Table 3 Parameter estimates for CAPRISA 002 AI study data across several quantiles

Parameter	$\widehat{Q}_{0.05}$ (SE)	$\widehat{Q}_{0.25}$ (SE)	$\widehat{Q}_{0.5}$ (SE)	$\widehat{Q}_{0.75}$ (SE)	$\widehat{Q}_{0.85}$ (SE)	$\widehat{Q}_{0.95}$ (SE)
Intercept	19.996 (1.161)*	22.171 (1.403)*	24.628 (1.464)*	26.595(1.419)*	27.972 (1.420)*	31.381 (1.397)*
Time	0.063 (0.015)*	0.069 (0.013)*	0.056 (0.013)*	0.046 (0.013)*	0.041 (0.013)*	0.034 (0.015)*
SQRT of time	− 0.866 (0.142)*	− 0.871 (0.129)*	− 0.695 (0.117)*	− 0.593 (0.119)*	− 0.581 (0.124)*	− 0.385 (0.162)*
Baseline BMI	0.056 (0.021)*	0.078 (0.024)*	0.082 (0.026)*	0.112 (0.032)*	0.131 (0.033)*	0.145 (0.030)*
Log of baseline VL	− 0.564 (0.078)*	− 0.568 (0.103)*	− 0.641 (0.096)*	− 0.713 (0.093)*	− 0.714 (0.089)*	− 0.739 (0.084)*
Post HAART initiation	1.683 (0.054)*	2.125 (0.073)*	2.560 (0.088)*	3.021 (0.096)*	3.114(0.097)*	2.287 (0.089)*
Age	0.021 (0.025)	0.029 (0.029)	0.029 (0.031)	0.029 (0.032)	0.026 (0.032)	0.013 (0.030)
Log-lik	− 18,454.68	− 17,169.85	− 16,828.96	− 17,344.63	− 17,952.50	− 19,088.77
AIC	36,937.36	34,367.69	33,685.92	34,717.25	35,933	38,205.55

* Significance at 5% level. See, Additional file 1, for more significant test results and confidence intervals

effect on the immune systems increases (Table 3). From low quantiles to upper quantiles, the post HAART initiation effect on CD4 cell counts has an increasing trend, and then at high quantile 0.95, its effect begins to decline.

R package *qrLMM()* sample outputs using CAPRISA 002 Acute Infection Study data across all fitted quantile levels can be found in Additional files 1, 2.

The results in graphical representation following QR-LMM over the framework of quantiles $\tau = \{0.05, 0.25, 0.5, 0.75, 0.85, 0.95\}$ are displayed in Fig. 1. The graph shows that the 95% confidence interval for the covariates effect and the nuisance parameter σ. The figure reveals that the effect of baseline BMI (β_4), and post HAART initiation (β_6) become more prominent across quantile levels, with their effect becoming more for higher conditional quantiles. Additionally, although the effects of time (β_2) and baseline VL (β_5) exhibit a significant positive and negative influence, respectively, on CD4 counts across all quantiles, the difference changes with regard to the lower quantiles. The $\widehat{\sigma}$ is symmetric about $\tau = 0.5$, taking its maximum value at that point and decreasing for higher quantiles. The convergence of estimates for all parameters was also evaluated using the graphical criteria (see Additional files 1, 2).

Conclusions

This study considered a quantile mixed-effects model with a likelihood-based function that adopts an ALD for the error term. We used the SAEM algorithm for determining exact ML estimates of the covariates effect and variance–covariance elements across a set of quantiles. We applied this methodology to the CAPRISA 002 AI Study data and illustrated how the procedure can be used to obtain robust parameters estimates when the interest is to get the estimation not only on the central location but also on the non-central locations of the conditional distribution, which brings a comprehensive and more complete picture of the effects. A series of QR-LMM at $\tau = 0.05, 0.25, 0.5, 0.75, 0.85$ and 0.95 were estimated (Table 3, and Additional files 1, 2), and the results were discussed.

Since quantile inference for discrete longitudinal data cannot thus be carried out directly yet, we modeled a continuous approximation form of the quantile function by using square-root-transformed CD4 count as the response variable. Time since seroconversion, HAART initiation, and baseline characteristics of the patients such as BMI, age, and VL was included in the study. It was found that except age, all the studied variables were found to have a significant effect on CD4 cell counts of HIV-infected patients across all quantiles. Although significant CD4 cell recovery in response to post HAART initiation across all quantiles was recognized, HIV-infected patients who were enrolled in the treatment with a high level of VL showed a significant negative effect on CD4 cell counts at upper quantiles [33]. Even though patients with higher BMI at baseline have improved CD4 cell count overtime after the treatment, they should not be ignored clinically. The study also suggested that physicians should carefully monitor patients with low BMI before and after the treatment because BMI can influence drug metabolism and, consequently, the immunological response to HAART [31, 33]. With the growing recognition of the quantile mixed-effects model, it looks practical that the methodology will be extended to a vast range of statistical applications such as binary data, multi-level models, survival analysis, and other areas of application, and these shall be the subject of future works.

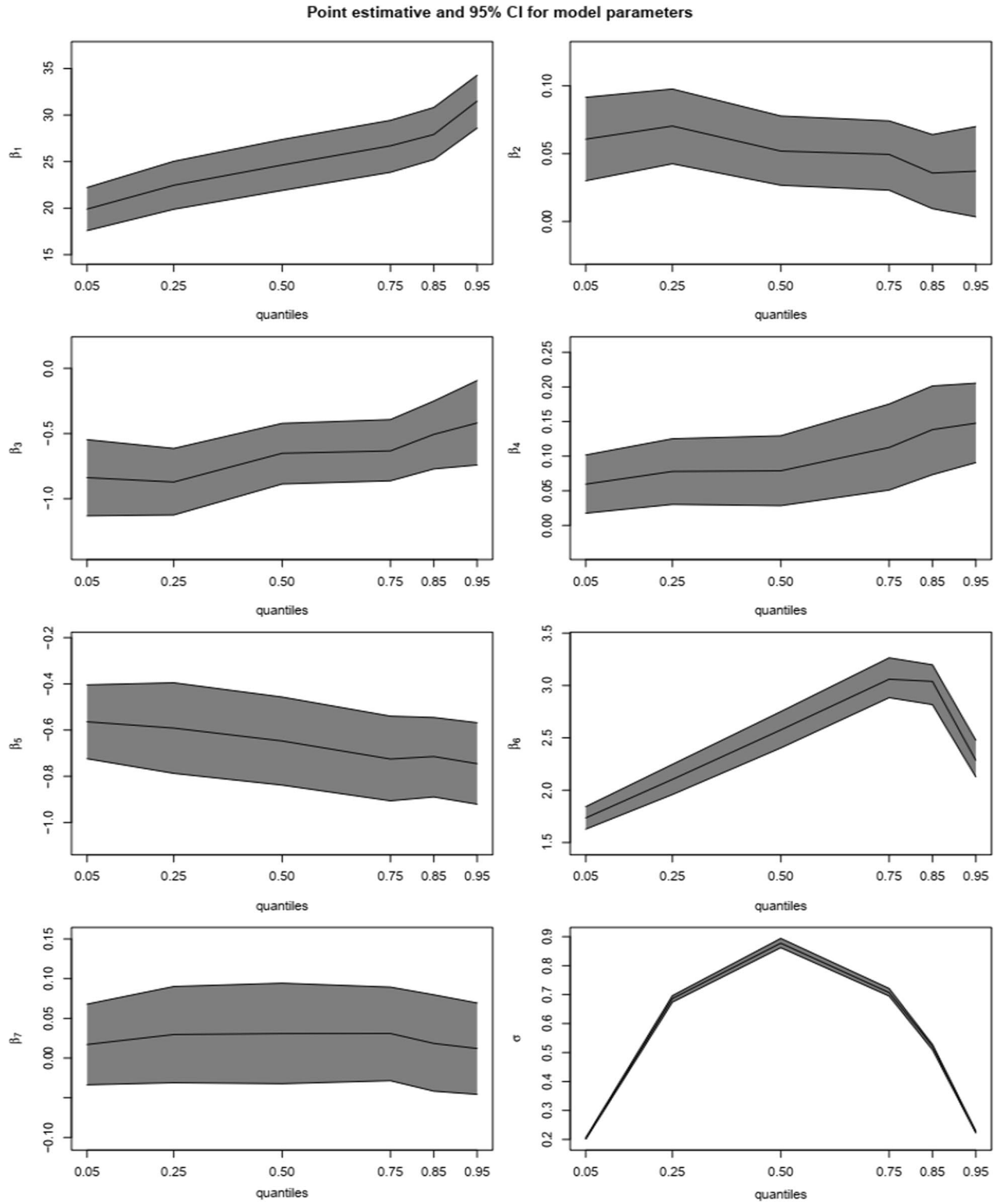


Fig. 1 Point estimates and 95% confidence bands for model parameters following the QR-LMM to the CAPRISA 002 AI Study data across various quantiles

Abbreviations
AI: Acute infection; AIDS: Acquired immune deficiency syndrome; ALD: Asymmetric Laplace distribution; ART: Antiretroviral therapy; ARV: Antiretroviral (drug); CAPRISA: Centre for the AIDS Programme of Research in South Africa; CD4: Cluster of difference 4 cell (T-lymphocyte cell); E-step: Expectation step; EM: Expectation–maximization; HAART: Highly Active Antiretroviral Therapy; HIV: Human immunodeficiency virus; LP: Linear programming; M-step: Maximization step; ML: Maximum likelihood; QR: Quantile regression; QR-LMM: Quantile regression for linear mixed-effects models; SAEM: Stochastic Approximation version of the EM algorithm; SE: Standard error; VL: Viral load refers to the number of HIV copies in a milliliter of blood (copies/ml).

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12879-021-06942-7.

Additional file 1. R package *qrLMM()* sample output using CAPRISA 002 Acute Infection Study data across fitted quantile levels.

Additional file 2. Graphic overview of convergence for model parameters across all fitted quantiles, produced from the *qrLMM* package using the CAPRISA 002 AI Study data.

Acknowledgements
We gratefully acknowledge CAPRISA for giving us access to the CAPRISA 002: Acute Infection Study data. CAPRISA is funded by the National Institute of Allergy and Infectious Diseases (NIAID), National Institutes for Health (NIH), and U.S. Department of Health and Human Services (Grant: AI51794). The authors would also like to thank Dr. Nonhlanhla Yende-Zuma (Head of Biostatistics unit at CAPRISA) for her cooperation, assistance, and technical support.

Authors' contributions
AAY obtained the data, did the analysis, and prepared the manuscript. AAY, SFM, HGM, and DGA planned the research problem. All authors deliberated on the results and consequences and commented on the paper at all stages. All authors contributed extensively to the work presented in this manuscript. All authors read and ratified the ultimate manuscript. All authors read and approved the final manuscript.

Funding
This work was supported through the DELTAS Africa Initiative and the University of KwaZulu-Natal. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [Grant 107754/Z/15/Z], DELTAS Africa Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB) programme and the UK government. The views expressed in this manuscript are those of the author(s). The funding body had no role in the design of the study and collection, analysis, and interpretation of data, and in writing the manuscript.

Availability of data and materials
The dataset used for this study can be obtained by requesting Dr. Nonhlanhla Yende-Zuma (Head of Biostatistics Unit, CAPRISA, Email: Nonhlanhla.Yende@caprisa.org) on reasonable request.

Declarations

Ethics approval and consent to participate
The study was approved by the Research Ethics Committee of the University of KwaZulu-Natal (E013/04), the University of the Witwatersrand (MM040202), and the University of Cape Town (025/2004). All participants provided written informed consent. All methods were performed following the relevant guidelines and regulations expressed in the Declaration of Helsinki.

Consent to publish
Not applicable.

Competing interests
The authors affirm that they have no competing interests, monetary or otherwise.

Author details
[1]School of Mathematics, Statistics, and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, Private Bag X01, Scottsville 3209, South Africa. [2]Institute of Human Virology, School of Medicine, University of Maryland, Baltimore, MD 21201, USA.

Received: 16 November 2020 Accepted: 3 December 2021
Published online: 04 January 2022

References

1. AIDSMAP. CD4 cell counts | aidsmap. Key points-May. 2017. https://www.aidsmap.com/about-hiv/cd4-cell-counts.
2. WHO. Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection: recommendations for a public health approach. 2016. https://apps.who.int/iris/bitstream/handle/10665/208825/9789241549684_eng.pdf. Accessed 24 Sept 2020.
3. Davino C, Furno M, Vistocco D. Quantile regression: theory and applications, vol. 988. Hoboken: Wiley; 2013.
4. Girma S, Görg H. Foreign direct investment, spillovers and absorptive capacity: evidence from quantile regressions. Bundesbank Series 1 Discussion Paper. 2005.
5. Chunying Z. A quantile regression analysis on the relations between foreign direct investment and technological innovation in China. In: 2011 international conference of information technology, computer engineering and management sciences, Vol. 4, IEEE. 2011. pp. 38–41.
6. Mirnezami R, Nicholson J, Darzi A. Preparing for precision medicine. N Engl J Med. 2012;366(6):489–91.
7. Koenker R, Bassett G Jr. Regression quantiles. Econometrica J Econom Soc. 1978;46(1):33–50.
8. Pinheiro J, Bates D. Mixed-effects models in S and S-PLUS. Berlin: Springer Science & Business Media; 2006.
9. Verbeke G, Molenberghs G. Linear mixed models for longitudinal data. Berlin: Springer Science & Business Media; 2009.
10. Twisk JW. Applied longitudinal data analysis for epidemiology: a practical guide. Cambridge: Cambridge University Press; 2013.
11. Diggle P, Diggle PJ, Heagerty P, Liang K-Y, Heagerty PJ, Zeger S. Analysis of longitudinal data. Oxford: Oxford University Press; 2002.
12. Brown H, Prescott R. Applied mixed models in medicine. Hoboken: Wiley; 2015.
13. Koenker R. Quantile regression. Cambridge: Cambridge University Press; 2005.
14. Wichitaksorn N, Choy SB, Gerlach R. A generalized class of skew distributions and associated robust quantile regression models. Can J Stat. 2014;42(4):579–96.
15. Galvao AF Jr. Quantile regression for dynamic panel data with fixed effects. J Econom. 2011;164(1):142–57.
16. Fu L, Wang Y-G. Quantile regression for longitudinal data with a working correlation model. Comput Stat Data Anal. 2012;56(8):2526–38.
17. Lipsitz SR, Fitzmaurice GM, Molenberghs G, Zhao LP. Quantile regression methods for longitudinal data with drop-outs: application to CD4 cell counts of patients infected with the human immunodeficiency virus. J R Stat Soc: Ser C (Appl Stat). 1997;46(4):463–76.
18. Galarza Morales CE. Quantile regression for mixed-effects models. 2015. https://bit.ly/3i7BPyQ.
19. Geraci M, Bottai M. Quantile regression for longitudinal data using the asymmetric Laplace distribution. Biostatistics. 2007;8(1):140–54.
20. Geraci M, Bottai M. Linear quantile mixed models. Stat Comput. 2014;24(3):461–79.
21. Galarza CE, Lachos VH, Bandyopadhyay D. Quantile regression in linear mixed models: a stochastic approximation EM approach. Stat Interface. 2017;10(3):471–82.
22. Reich BJ, et al. Flexible Bayesian quantile regression for independent and clustered data. Biostatistics. 2010;11(2):337–52.

23. Noufaily A, Jones M. Parametric quantile regression based on the generalized gamma distribution. J R Stat Soc: Ser C (Appl Stat). 2013;62(5):723–40.
24. Liu Y, Bottai M. Mixed-effects models for conditional quantiles with longitudinal data. Int J Biostat. 2009;5(1): 28.
25. Muir PR, Wallace CC, Done T, Aguirre JD. Limited scope for latitudinal extension of reef corals. Science. 2015;348(6239):1135–8.
26. Fornaroli R, Cabrini R, Sartori L, Marazzi F, Vracevic D, Mezzanotte V, Annala M, Canobbio S. Predicting the constraint effect of environmental characteristics on macroinvertebrate density and diversity using quantile regression mixed model. Hydrobiologia. 2015;742(1):153–67.
27. Blankenberg S, Salomaa V, Makarova N, Ojeda F, Wild P, Lackner KJ, Jørgensen T, Thorand B, Peters A, Nauck M. Troponin I and cardiovascular risk prediction in the general population: the BiomarCaRE Consortium. Eur Heart J. 2016;37(30):2428–37.
28. Patel DE, Geraci M, Cortina-Borja M. Modeling normative kinetic perimetry isopters using mixed-effects quantile regression. J Vis. 2016;16(6):7–7.
29. Van Loggerenberg F, Mlisana K, Williamson C, Auld SC, Morris L, Gray CM, Karim QA, Grobler A, Barnabas N, Iriogbe I. Establishing a cohort at high risk of HIV infection in South Africa: challenges and experiences of the CAPRISA 002 Acute Infection Study. PLoS ONE. 2008;3(4):e1954.
30. Mlisana K, Werner L, Garrett NJ, McKinnon LR, van Loggerenberg F, Passmore J-AS, Gray CM, Morris L, Williamson C, Abdool Karim SS. Rapid disease progression in HIV-1 subtype C-infected South African Women. Clin Infect Dis. 2014;59(9):1322–31.
31. Yirga AA, Melesse SF, Mwambi HG, Ayele DG. Modelling CD4 counts before and after HAART for HIV infected patients in KwaZulu-Natal South Africa. Afr Health Sci. 2020;20(4):1546–61.
32. Yirga AA, Melesse SF, Mwambi HG, Ayele DG. Negative binomial mixed models for analyzing longitudinal CD4 count data. Sci Rep. 2020;10(1):1–15.
33. Yirga AA, Melesse SF, Mwambi HG, Ayele DG. Additive quantile mixed effects modelling with application to longitudinal CD4 count data. Sci Rep. 2021;11(1):1–12.
34. Whelan D. Gender and HIV/AIDS: taking stock of research and programmes. Geneva: UNAIDS; 1999.
35. UN Women, 2014. Message from UN Women's Executive Director for World AIDS Day, 1 December 2014. https://www.unwomen.org/en/news/stories/2014/12/world-aids-day-2014.
36. amfAR. The Foundation for AIDS Research. Statistics: women and HIV/AIDS. 2015. https://www.amfar.org/about-hiv-and-aids/facts-and-stats/statistics--women-and-hiv-aids/.
37. Kassutto S, Rosenberg ES. Primary HIV type 1 infection. Clin Infect Dis. 2004;38(10):1447–53.
38. Cohen MS, Shaw GM, McMichael AJ, Haynes BF. Acute HIV-1 infection. N Engl J Med. 2011;364(20):1943–54.
39. Rosenberg ES, Altfeld M, Poon SH, Phillips MN, Wilkes BM, Eldridge RL, Robbins GK, Richard T, Goulder PJ, Walker BD. Immune control of HIV-1 after early treatment of acute infection. Nature. 2000;407(6803):523–6.
40. Yirga AA, Ayele DG, Melesse SF. Application of quantile regression: modeling body mass Index in Ethiopia. Open Public Health J. 2018;11(1):221–33.
41. Buchinsky M. Recent advances in quantile regression models: a practical guideline for empirical research. J Hum Resour. 1998;33(1):88–126.
42. Ellerbe CN, Gebregziabher M, Korte JE, Mauldin J, Hunt KJ. Quantifying the impact of gestational diabetes mellitus, maternal weight and race on birthweight via quantile regression. PLoS ONE. 2013;8(6):e65017.
43. Koenker R, Hallock KF. Quantile regression. J Econ Perspect. 2001;15(4):143–56.
44. Peterson MD, Krishnan C. Growth charts for muscular strength capacity with quantile regression. Am J Prev Med. 2015;49(6):935–8.
45. Song X, Li G, Zhou Z, Wang X, Ionita-Laza I, Wei Y. QRank: a novel quantile regression tool for eQTL discovery. Bioinformatics. 2017;33(14):2123–30.
46. Sherwood B, Wang L, Zhou XH. Weighted quantile regression for analyzing health care cost data with missing covariates. Stat Med. 2013;32(28):4967–79.
47. Cook BL, Manning WG. Measuring racial/ethnic disparities across the distribution of health care expenditures. Health Serv Res. 2009;44(5p1):1603–21.
48. Borgoni R. A quantile regression approach to evaluate factors influencing residential indoor radon concentration. Environ Model Assess. 2011;16(3):239–50.
49. Yu K, Lu Z, Stander J. Quantile regression: Applications and current research areas. J R Stat Soc: Ser D (The Statistician). 2003;52(3):331–50.
50. Knight CA, Ackerly DD. Variation in nuclear DNA content across environmental gradients: a quantile regression analysis. Ecol Lett. 2002;5(1):66–76.
51. Cade BS, Noon BR. A gentle introduction to quantile regression for ecologists. Front Ecol Environ. 2003;1(8):412–20.
52. Lachos VH, Chen M-H, Abanto-Valle CA, Azevedo CL. Quantile regression for censored mixed-effects models with applications to HIV studies. Stat Interface. 2015;8(2):203.
53. Koenker R, Machado JA. Goodness of fit and related inference processes for quantile regression. J Am Stat Assoc. 1999;94(448):1296–310.
54. Cameron AC, Trivedi PK. Microeconometrics: methods and applications. Cambridge: Cambridge University Press; 2005.
55. Cameron AC, Trivedi PK. Microeconometrics using stata, vol. 2. College Station: Stata Press; 2009.
56. Cameron AC, Trivedi PK. Regression analysis of count data, vol. 53. Cambridge: Cambridge University Press; 2013.
57. Yu K, Moyeed RA. Bayesian quantile regression. Stat Probab Lett. 2001;54(4):437–47.
58. Yu K, Zhang J. A three-parameter asymmetric Laplace distribution and its extension. Commun Stat Theory Methods. 2005;34(9–10):1867–79.
59. Kotz S, Kozubowski T, Podgorski K. The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance. Berlin: Springer Science & Business Media; 2012.
60. Kozubowski TJ, Nadarajah S. Multitude of Laplace distributions. Stat Pap. 2010;51(1):127.
61. Winkelmann R. Econometric analysis of count data. Berlin: Springer Science & Business Media; 2008.
62. Hilbe JM. Negative binomial regression. Cambridge: Cambridge University Press; 2011.
63. Hilbe JM. Modeling count data. Cambridge: Cambridge University Press; 2014.
64. Machado JAF, Silva JS. Quantiles for counts. J Am Stat Assoc. 2005;100(472):1226–37.
65. Winkelmann R. Reforming health care: evidence from quantile regressions for counts. J Health Econ. 2006;25(1):131–45.
66. Miranda A. Planned fertility and family background: a quantile regression for counts analysis. J Popul Econ. 2008;21(1):67–81.
67. Dempster AP, et al. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc: Ser B (Methodol). 1977;39(1):1–22.
68. McLachlan GJ, Krishnan T. The EM algorithm and extensions, vol. 382. Hoboken: Wiley; 2007.
69. Delyon B, Lavielle M, Moulines E. Convergence of a stochastic approximation version of the EM algorithm. Ann Stat. 1999;21(1):94–128.
70. Jank W. Implementing and diagnosing the stochastic approximation EM algorithm. J Comput Graph Stat. 2006;15(4):803–29.
71. Meza C, Osorio F, De la Cruz R. Estimation in nonlinear mixed-effects models using heavy-tailed distributions. Stat Comput. 2012;22(1):121–39.
72. Kuhn E, Lavielle M. Coupling a stochastic approximation version of EM with an MCMC procedure. ESAIM Probab Stat. 2004;8:115–31.
73. Kuhn E, Lavielle M. Maximum likelihood estimation in nonlinear mixed effects models. Comput Stat Data Anal. 2005;49(4):1020–38.

Publisher's Note