
COLLEGE OF NATURAL SCIENCE

DEPARTMENT OF STATISTICS

# A Joint Modeling Approach for Analysis of Longitudinal Body Weight and Sputum Status of Tuberculosis patients in Jimma University Specialized Hospital

By:

Mersha Filate

A Thesis submitted to Department of Statistics, School of Graduate Studies, College of Natural Science, Jimma University, in Partial Fulfillment for the Requirement of Masters of Science in Biostatistics

September, 2014
Jimma, Ethiopia

# A Joint Modeling Approach for Analysis of Longitudinal Body Weight and Sputum Status of Tuberculosis patients in Jimma University Specialized Hospital

By:

Mersha Filate

Advisor: Wondwosen Kassahun (Ph.D)

Co.advisor: Girma Tefera (Msc.)

September, 2014
Jimma, Ethiopia

## Table of Contents

Content page

Declaration ..... i

Dedication ..... ii

Acknowledgment ..... iii

Abstract ..... iv

Acronyms ..... v

List of Tables ..... vi

List of Figures ..... vii

INTRODUCTION ..... 1

1.1. Background ..... 1

1.2. Longitudinal Versus Cross -Sectional Studies ..... 2

1.3. Modeling Longitudinal Outcomes ..... 3

1.4. Joint Modeling ..... 4

1.5. Statements of the Problem ..... 6

1.6. Objectives ..... 7

1.6.1. General Objective ..... 7

1.6.2. Specific Objectives ..... 7

1.7. Significance of the Study ..... 7

LITERATURE REVIEW ..... 8

2.1. Risk Factors and Related Study ..... 8

2.2. TB Treatment Regimen ..... 10

2.3. Overview of Models for Longitudinal Outcomes ..... 11

2.3.1. Linear Mixed Model ..... 11

2.3.2. Generalized Linear Mixed Model ..... 11

2.3.3. Joint Modeling Approach ..... 12

METHODS ..... 14
3.1. Study Population and Design ..... 14
3.2. Data Source and Description ..... 14
3.3. Variables of Interest ..... 15
3.3.1. Dependent Variable ..... 15
3.3.2. Predictor Variables (Independent variable) ..... 15
3.4. Exploratory Data Analysis ..... 16
3.4.1. Individual Profile Plot ..... 16
3.4.2. The Average Evolution ..... 17
3.4.3. The Variance Structure ..... 17
3.4.4. The Correlation Structure ..... 17
3.5. Statistical Models ..... 17
3.5.1. Models for a Single Longitudinal Continuous Response ..... 17
3.5.1.1. Linear Mixed Model ..... 17
3.5.1.1.1. Covariance Structure of Linear Mixed Model ..... 18
3.5.2. Generalized Linear Model ..... 20
3.5.2.1. Models for a Single Longitudinal Binary Response ..... 21
3.5.2.1.1. The Generalized Linear Mixed Model (GLMM) ..... 21
3.5.3. Joint Model for Continuous and Binary Responses ..... 22
3.5.4. Parameter Estimation Methods ..... 25
3.5.4.1. Parameter Estimation of LMM ..... 25
3.5.4.1.1. Maximum Likelihood Estimation ..... 25
3.5.4.1.2. Restricted Maximum Likelihood Estimation ..... 26
3.5.4.2. Parameter Estimation of GLMM ..... 27
3.5.4.3. Parameter Estimation of the Joint Model ..... 27

3.5.5. Model Comparison Technique ..... 28
3.5.6. Model Diagnosis ..... 29
3.6. Ethical Consideration ..... 29
ANALYSIS AND RESULTS ..... 30
4.1. Baseline Information ..... 30
4.2. Separate Analysis of Continuous Longitudinal Outcome (Weight) ..... 32
4.2.1. Exploratory Analysis ..... 32
4.2.1.1. Individual profiles plot of Body Weight ..... 32
4.2.1.2. Exploring Mean Structure of Body Weight of Patients' ..... 34
4.2.1.3. Exploring the Variability of Weight of TB patients' ..... 36
4.2.1.4. Exploring the Correlation Structure ..... 39
4.2.2. Separate Linear Mixed Model for Body Weight ..... 40
4.2.2.1. Random Effect Selection ..... 40
4.2.2.2. Linear Mixed Model for Body Weight with Linear Time Effect ..... 41
4.2.2.3. Pattern of Variance-Covariance Structure ..... 43
4.3. Separate Analysis of Binary Longitudinal Outcome (Sputum Conversion) ..... 45
4.3.1. Generalized Linear Mixed Model for Sputum Conversion ..... 45
4.4.Joint Model of Weight and Sputum Conversion ..... 47
4.4.1. Comparison of Joint and Separate Models ..... 50
4.5. Discussion ..... 54
CONCLUSION AND RECOMMENDATION ..... 57
5.1. Conclusion ..... 57
5.2. Recommendation ..... 57
References ..... 59
Appendix -1: Model diagnosis for Linear Mixed Model ..... 64
Appendex-2: Model diagnosis for Generalized Linear Mixed Model ..... 65

JIMMA UNIVERSITY

COLLEGE OF NATURAL SCIENCE

DEPARTMENT OF STATISTICS

A Joint Modeling Approach for Analysis of Longitudinal Body Weight and Sputum Status of Tuberculosis patients in Jimma University Specialized Hospital

BY

Mersha Filate

As members of the Board of Examiners of M.Sc. thesis open defense examination of the above title, we read and evaluated the thesis and examined the candidate

| | | |
|---|---|---|
| Name of Chairman | Signature | Date |
| Name of Major advisor | Signature | Date |
| Name of Co-advisor | Signature | Date |
| Name of Internal Examiner | Signature | Date |
| Name of Examiner | Signature | Date |

# Declaration

I declare that this thesis is a result of my genuine work and all sources of materials used, for writing it, have been duly acknowledged. I have submitted this thesis to Jimma University in the partial fulfillment for the Degree of Master of Science in Biostatistics. The thesis can be deposited in the university library to be made available to borrowers for reference. I solemnly declare that I have not so far submitted this thesis to any other institution anywhere for that award of any academic degree, diploma or certificate.

Brief quotations from this thesis are allowed without requiring special permission provided that an accurate acknowledgement of the source is made. Requisites for extended quotations for the reproduction of the thesis in whole or in part may be granted by the head of the department of statistics when in his or her judgment the proposed use of the material is for a scholarly interest. In all other instances, however, permission must be obtained from the author.

Mersha Filate

Date: _____

Signature:_____

Jimma University, Jimma, Ethiopia

.

# Dedication

This study is dedicated to my mother Abebu Tarekegn and to my brother Gedefaw Filate.It also dedicated to my friends those who were with me.

## Acknowledgment

First and foremost, praise and thanks must be given the Almighty God, the creator and the guardian, for giving me the courage, patience, strength and the knowledge to complete this thesis successfully.

Next, I would like to express my sincerely appreciation to my advisor Wondwosen Kassahun (Ph.D) for his insight, wise direction, constant encouragement, and for his support in so many aspects to the successful realization of the thesis.

I am thankful to my co-advisor Mr. Girma T. (M.Sc.) for his valuable comments, material and moral supports.

I also express my heartfelt thanks and gratitude to my friends Mr.Zelalem M. (M.Sc.) and Mr.Sisay W. (M.Sc.) for their kindly encouragement and support throughout the study period.

Finally, I am so indebted to a supportive families especially my mother Abebu Tarekegn and my brother Mr. Gedefaw F. and my friends those who were with me.

# Abstract

**Background**:-Tuberculosis is a major public health problem even though it is treatable and curable. Weight and sputum conversion during anti tuberculosis (TB) treatment period is an important component and they have been described as a useful marker to assess the progress of TB patients'.

**Objective**:-The objective of this study is to fit a joint model in which both the longitudinal weight and sputum status are studied to investigate their joint evolution and identify the risk factors for the body weight and sputum status of tuberculosis patients in Jimma University specialized Hospital during six months diagnosis period.

**Method:** The data for this thesis were obtained from a retrospective study from TB patients registered between 2011 and 2013. The following statistical models were considered: linear mixed model for the separate body weight analysis, generalized linear mixed model for sputum status and a joint model with correlated random effects was fitted to simultaneously study the evolution over time of a longitudinal body Weight and Sputum status. The estimation of the model parameters was done by maximum and restricted likelihood and maximum likelihood based on adaptive Gaussian hermite Quadrature as implemented in the SAS procedure NLMXED.

**Result:** The overall proportion of tuberculosis patients during follow up time having positive and negative sputum status is 39.3% and 60.7% respectively. Based on the data exploration the mean change of body weight has a linear relation with time. From the separate linear mixed model all covariates (types of TB, age, dose) are significant and their interaction by time were the risk factors for the body weight of TB patients. In case of separate generalized linear mixed model age, types of TB, dose and time have a significant effect on the sputum status of TB patients. Similar covariates were significant in the joint model of body weight and sputum status and estimates were found to be very close to separate analysis. But, the joint model yields higher precision and allows for quantifying the association between outcomes and association between the outcomes in this joint model was negative ($\rho$ =-0.698, p=0.0001).

**Conclusion**: The results of the separate and joint models almost the same. When the joint model is compared with the separate model, it is both the most parsimonious model and also fits the data better than the separate model. The joint model showed that the body weight and positive sputum status are inversely related each other.

# Acronyms

| | |
|---|---|
| AFB | Acid-Fast Bacilli |
| AIDS | Acquired Immunodeficiency Syndrome |
| AIC | Akaike's information criterion |
| ARV | Antiretroviral medicine |
| DOT | Directly Observed Treatment |
| DOTS | Directly Observed Treatment, Short Course |
| DR-TB | Drug- resistant tuberculosis |
| GLM | Generalized Linear Model |
| GLMM | Generalized Linear Mixed Model |
| HIV | Human immunodeficiency virus |
| MDR-TB | Multi-Drug Resistant Tuberculosis |
| ML | Maximum likelihood |
| REML | Restricted Maximum likelihood |
| M. tuberculosis | Mycobacterium tuberculosis |
| LMM | linear mixed model |
| PTB | Pulmonary Tuberculosis |
| RNTCP | Revised National TB Control Programme |
| TB | Tuberculosis |
| HIV | Human Immunodeficiency virus |
| WHO | World Health Organization |

# List of Tables

Table 1: Description of predicted variable included in the model…….....16

Table 2.The mean and standard deviation of weight over time according to the corresponding covariates ……...……...……..……31

Table3. Proportion of sputum conversion with baseline Categorical Covariates.....31

Table 4: The correlation matrix of the weight of TB patients…….....39

Table5: Selection of random effect for continuous longitudinal data weight……..41

Table6. Comparison of linear time effect model with quadratic time effects…….....41

Table7. Parameter estimates and standard errors for the separate LMM for body weight of the final model using ML and REML…….....43

Table 8. Comparison of model with different correlation function for weight ……..….44

Table 9.Standard errors and covariance structure for random effects (LMM)……...…….45

Table 10: Parameter estimates and standard errors for GLMM…….…..……46

Table 11.A parameter estimate and standard errors of the joint model for body weight and sputum conversion……..…….....48

Table 12: The Variance-Covariance estimates for the joint model…….…....49

Table 13: The correlation matrix of the Random effect in the joint model…….49

Table 14: Parameter estimates and standard errors for separate and joint model……..….51

Table 15: Covariance estimates and correlation estimates for separate and joint model…...53

# List of Figures

Figure 4.1. Individual plot of weight over time……...32

Figure 4.2. Individual plot of weight by sex ……...33

Figure 4.3. Individual plot of weight by category of TB ……33

Figure 4.4. Individual plot by Sex: HIVstatus ……..33

Figure 4.5.Individual plot of weight by sex: HIVstatus……..33

Figer 4.6.Individual plot of the weight of TB patient by HIV status and Types of TB……...34

Figer 4.7.The mean profile plot of the weight of TB patient over time……..…….....34

Figer 4.8.The mean profile plot of the weight of TB patient over time by sex…….....34

Figer 4.9.The mean profile plot of the weight over time by TB category……...…..34

Figure 4.10.The mean profile plot by HIV status ……...35

Figure 4.11.The mean profile plot by Sex and category of TB…….....…35

Figure 4.12.The mean profile plot by Sex and HIV status…….……..…...35

Figure 4.13.The mean profile plot by category of TB and HIV status……...….36

Figure4.14. The variance profile plot of the weight of TB patient over time…….....37

Figure4.15. The variance profile plot of the weight of TB patient by sex…….....37

Figure4.16. The variance plot by HIV status……...…….....37

Figer 4.17.The variance plot of body weight by TB category……..……...38

Figer 4.18.The variance plot of body weight by Sex: TB category…….....38

Figure 4.19.The variance plot by sex and HIV status …….……....38

Figure 4.20.The variance plot by HIV status and TB category……...38

Figer 4.21.Interval plot for subject specific intercept and slope .....39

Figer 4.22.Scatter plot matrix for the weight of TB patient……....40

**CHAPTER-ONE**

# INTRODUCTION

## 1.1. Background

Tuberculosis is an infectious disease caused by various strains of mycobacteria, especially Mycobacterium tuberculosis and usually attacks the lung (Smith I., 2003).It remains to be a major cause of morbidity and mortality throughout the world. It is estimated that one-third of the world's population is infected, 8.8 million people develop TB, and 1.45 million people die annually from the disease (WHO, 2011). In Africa, about 2.8 million incident TB cases and 390 thousands TB deaths occurred in 2009 (WHO, 2010). Ethiopia ranks seventh among the world's 22 high-burden TB countries (WHO, 2011). According to the World Health Organization's (WHO) Global TB Report 2011, Ethiopia had an estimated incidence rate of 261 cases per 100,000 population and 29 thousands deaths in 2010, with an estimated prevalence rate of 394 cases per 100,000 populations (WHO, 2010).

Tuberculosis can affect any organ system in the body, this infection may also manifest in other parts of the body including the spinal cord, kidneys or brain. Symptoms of an active tuberculosis infection include exhaustion, fever, nausea, chest pain and the presence of blood in urine or as a result of persistent coughing. Weight loss may also occur as a side effect of the previous symptoms or as a separate one of its own. TB is a wasting disease and bodyweight variation has been proposed as a practical anthropometric marker to predict TB treatment outcome .Moreover, weight loss of 2 kg. or more during the first-month therapy has been considered as a potential risk factor for toxicity due to drugs. Many countries, including Peru, routinely weigh patients and repeat sputum microscopy tests on a monthly basis during therapy to assess treatment response. Several studies have reported that positive sputum microscopy at second month of treatment is associated with subsequent treatment failure, but is insensitive at population level .Thus, patients' bodyweight might be a helpful and cheap test to predict TB treatment outcome (Becerra MC.et al., 2000). Positive sputum conversion has serious consequences, including ongoing infectivity and development of drug-resistant Mycobacterium tuberculosis. No reliable way exists to predict which patient will complete TB treatment; however, failure to complete treatment has been associated with alcohol abuse, drug abuse, and homelessness (Brudney K. and Dobkin J., 1991).

Also, patients with AIDS have been found to be more likely than those without AIDS to complete treatment (Brudney K, and Dobkin J. (1991)).The extent to which other factors, including program quality, influence the outcome of treatment has not been explored.

## 1.2. Longitudinal Versus Cross -Sectional Studies

Longitudinal data require that subjects in the study be repeatedly measured across time (Diggle.et.al, 2002; Hedeker & Gibbons, 2006; Vonesh & Chinchilli, 1997). This is the crucial difference between longitudinal data and cross-sectional data, which measures only a single outcome for each individual. An advantage of longitudinal studies is having more information on each subject. With this extra information, researchers are able to observe a trajectory for the subjects. Individual trajectories show how the response variable changes over time for the respective individual. In gathering trajectories for all subjects, an overall trend and its relationship to covariates of interest may then be assessed. Cross-sectional data does not allow for distinguishing these changes over time within individuals (Diggle et al., 2002). More elegantly stated, repeated measurements from the same subject provide more independent information than a single measurement from a single subject as in cross-sectional studies (Hedeker & Gibbons, 2006).

For this reason, longitudinal studies are more powerful than cross-sectional studies (Hedeker & Gibbons, 2006). Often, the goal of longitudinal analysis is to investigate the effects of covariates both on the overall level of the response (outcome) and on changes of the response over time (Skrondal & Rabe-Hesketh, 2008).

Another characteristic of longitudinal data are that the data are clustered or considered two-level data (Skrondal & Rabe-Hesketh, 2008). In other words, values or measurements are nested within the individual as measurements are obtained at different time points. In general, individuals are considered at level 2 and the repeated observations within individuals are at level 1. Higher levels may exist beyond the individual level, but are not the focus of this thesis. Longitudinal data are a special case of multilevel or hierarchical data in that the measurements are in chronological order and consist of a large number of small clusters (Skrondal & Rabe-Hesketh, 2008). Longitudinal data are also characterized by missing (unbalanced) data and time-dependent covariates (Davis, 2002).Clustered observations from

the same subject are likely correlated. This correlation implies a violation of the independent observations assumption from traditional statistical methods and must be accommodated. Some consequences of ignoring the correlation include incorrect inferences about regression coefficients, inefficient and less precise estimates, and less protection against biases due to missing data (Diggle et al., 2002).

The outcome measured in longitudinal data may be continuous, binary, ordinal, or categorical in nature. Longitudinal data may be collected prospectively or retrospectively; prospective data, as in clinical trials, are typically preferred to minimize recollection bias (Diggle et al., 2002). Longitudinal studies may be applied to social sciences such as psychology and economics as well as the biological sciences and clinical trials for evaluating new drugs (Diggle et al., 2002; Vonesh & Chinchilli, 1997). Multilevel modeling has become increasingly popular, particularly in the area of education (Singer, 1998). For more examples of uses of longitudinal data outside of this thesis, please refer to Diggle et al. (2002) and Vonesh and Chinchilli (1997).

## 1.3. Modeling Longitudinal Outcomes

Longitudinal outcomes are a series of measurements of the same event taken from the same individual repeatedly over time. The most unique characteristic of longitudinal data is the ability to directly study change. The primary goal of most longitudinal studies is to characterize the change in response over time and the factors that influence this change.

Great strides have been made over the past three decades involving development of statistical methodology for longitudinal data analysis. Longitudinal data require special methodology because the series of data from one subject are likely intercorrelated, and this correlation must be taken into account to draw valid statistical inferences. In fact, longitudinal data usually exhibit a positive correlation, with the strength of the association decreasing as a function of time separation (i.e. observations further apart as less correlated than those closer together).The two most commonly used approaches to analyzing longitudinal data are referred to as marginal models (population-averaged) and random-effects (subject-specific) models. The marginal model describes the relationship between the outcome variable and explanatory variables with a population average regression, as in a cross-sectional study

(Diggle et al., 2002). This approach is sometimes called the population-averaged model as it attempts to reduce the repeated values to a summary statistic such as the mean or population average which includes Generalized Estimating Equation (GEE) and Marginalized Multilevel Model (( Heagerty, P. J. and Zeger, S. L. (2000)). This approach is not as practical in the presence of time varying covariates (Diggle et al., 2002). As previously mentioned, the repeated measurements are likely correlated since they are obtained from the same subject. To account for within-subject correlation in the marginal model, the mean and covariance are modeled separately (Diggle et al., 2002). Parameter estimates for population-averaged models depend on the degree of heterogeneity in the population and this may vary between populations (Skrondal & Rabe-Hesketh, 2008).

The random-effects (linear Mixed and Generalized linear Mixed) models, on the other hand, consider that regression coefficients vary across individuals (Diggle et al., 2002); a process that stems from the assumption that repeated observations are correlated. In basic terms, there is an average regression coefficient from which each individual deviates given person-specific conditions. The random-effects model is interested in how much each individual deviates from these common regression coefficients. Also of interest is how subjects vary between each other and how measurements for each subject vary. These deviations are often referred to as between-subject variations and within-subject variations. The random-effects model takes care of both. Hence, it is possible to estimate individual-level and population-level growth curve parameters. The approaches were depending upon the research question and objective of the study. This approach was focused with applications of body weight and sputum conversion of TB patients.

## 1.4. Joint Modeling

Joint modeling has received massive attention in recent years, owing to researchers' desire for more insight into their data with a single statistical model. The reason to find this type of analysis is because commonly researchers simultaneously record several kinds of outcomes in their studies. These outcomes are often of a mixed nature. Prevalent examples are situations where a combination of continuous, binary, ordinal, survival and missing outcomes occurs. Continuous and binary outcomes often appear in longitudinal studies where one observes follow up measurements on patients. Conducting a joint analysis allows addressing

additional scientifically relevant questions. For example, when one is interested in knowing whether a new treatment could improve all outcomes simultaneously or in the measurement of the association between the various responses and how this association evolves over time, a joint model is advisable. Also, joint models are popular owing to the fact that they ensure unbiased statistical inferences (minimization of variation of estimates) (Tsiatis et al., 1995; Wulfsohn and Tsiatis 1997) in a variety of settings.

In this thesis, a model has been built for a longitudinal binary process (sputum status) and a longitudinal continuous process (weight). The primary interest is in the setting of two processes: The model has been applied to see the relation between these two outcomes. The generalized linear mixed model component in a shared-parameter model and its so-called hierarchical extensions was replaced by the model of Heagerty (1999). A brief review was offered for correlated continuous and binary data. Full maximum likelihood estimation with iterative numerical Quadrature methods is adopted to obtain parameter estimates.

## 1.5. Statements of the Problem

Patients with Tuberculosis (TB) often suffer from severe weight loss, a symptom that is considered immune-suppressive and a major determinant of severity and disease outcome Van Crevel R.et al.(2002). The association between body weight, TB mortality and morbidity has been studied extensively since 1986 (England A.et al. (2003). Directly Observed Treatment Short-course (DOTS) is the internationally recommended strategy for TB control, adopted as the Revised National TB Control Programme (RNTCP) in India since 1997. The country was covered under the programme by March 2006 and has almost achieved the global target of 85% cure and 70% case detection. There are about 8.9 million patients with TB in India, of whom half are infectious (sputum positive (TB India 2005). Currently, nationwide coverage results in a success rate of 86% and a death rate of 4 % **(**http://www.tbcindia.org (Accessed on May April 2006).

The weight and sputum status of the patient taken at different time points during treatment are an important components to assess the progress of patients. The relationship between change in weight and sputum status among patients during anti-TB treatment and other factors such as socio-economic demographic characteristics, smoking and drinking habits, whether the patient took treatment under supervision, the type of DOT centers and problems in taking drugs has not been well documented. Although many papers have reported bodyweight as a marker to predict therapy failure, death or relapse, to our knowledge, no study has reported an appropriate joint longitudinal analysis of patients during TB treatment assessing bodyweight change over time and it association with sputum status in various applications, it is common to observe statistical problems with outcomes of a mixed nature as in Molenberghs and Verbeke (2005). The reasons why this study was conducted are:

- What factors influenced sputum status and body weight of TB patients?
- How the average body weight for TB patient changes over time?
- Does a covariate predict similar change in the given outcomes?
- How does the association between weight and sputum status evolve over time?
- How to evaluate the joint and independent effects of a set of predictors on a set of outcomes?

To dig out these research questions and also identify the risk factors related to the weight loss of TB patients, the main focus is modeling a joint modeling for two response variables.

## 1.6. Objectives

### 1.6.1. General Objective

The general objective of the study is to build a joint model in which both the longitudinal body weight and sputum status are associated through unobserved correlated random effects and identify the risk factors affecting the two end points.

### 1.6.2. Specific Objectives

The specific objectives of the study which have accomplished to achieve the general objective stated above are the following.

- To explore the mean evolution of weight of TB patients.
- To evaluate and assess change and trends of patients' bodyweight over time depending on TB treatment outcome
- To fit a separate model for body weight and sputum status and to identify the associated factors for weight and sputum status of TB patients.
- To fit a joint model for weight and sputum status that yields biologically as well as statistically plausible and interpretable estimates of the effect of important covariates on body weight and sputum status of TB patients.

## 1.7. Significance of the Study

The results of this study might have the following benefits:

- To government and other concerned bodies in setting policies, strategies and further investigation for reducing morbidity and death of TB patients'.
- To help donors and government to understand risk factors that influences the death of TB patients.
- It will serve as input for upcoming similar researches.
- To assessing the quality TB care service to determine whether standards are being practiced in private and government health facility.

# CHAPTER TWO

# LITERATURE REVIEW

## 2.1. Risk Factors and Related Study

TB is primarily a disease of the respiratory system which spreads when the TB patients expel the droplets by sneezing, spitting and coughing and the people nearby inhale the droplets and become infected with mycobacterium, mainly. When mycobacteria reach the alveoli of the lung, they invade and replicate within the endosomes of alveolar macrophages. Infection can result in latent TB or active disease which clinically can be classified as pulmonary-smear negative or pulmonary TB. Latent TB is asymptomatic and Symptoms for active TB are chronic cough, blood-tinged sputum, night sweats, and weight loss (Compoux JJ. et al., 2004).

A study conducted in Pampas de San Juan de Miraflores, a periurban shantytown among 530 patients started tuberculosis treatment during the period of study and were eligible for this study; but, 20 moved away before starting treatment, 37 abandoned therapy, 11 had previous failures, and 2 had unknown outcomes. Therefore, 460 (87%) patients were included in the analysis, 55.4% of them were males and the mean age was 31.6 years (SD: 14.1; range: 18–80). Of the total, 42 (9.1%) had a poor outcome at the end of tuberculosis therapy (17 sputum negative and 25 sputum positive).The data were fitted using Generalized estimating Equation, but, there was no significant difference between weights of outcome groups at baseline ($p = 0.12$); however, on average, weight decreased in those who developed an adverse outcome whereas it increased among those who ended treatment as cured (smear negative result or good outcome). When assessing correlation structure for repeated measurements using Quasi-information criterion (QIC), the best working correlation was exchangeable. Other structures (auto-regressive, unstructured, and non-stationary) were evaluated with the model, but they did not achieve convergence. In any case, robust standard errors were used to handle misspecification of variance or correlation functions (Hardin J and Hilbe JM, 2003).

Rios J.et al. (2011) examined the relationship between the body weights with sputum conversion in patients with tuberculosis. Results obtained from the marginal models, the coefficient for

adverse outcome was significant (p = 0.007), indicating that the difference in weight (about 2 kg) among patients with sputum positive and negative at baseline was statistically different. Similarly, the interaction terms together were significant (p = 0.002) indicating that changes of weight over time among patients with sputum positive differed of those with sputum negative. On the other hand, patients with poor outcome lost about 1 kg (0.97 kg according to the model) at the first month of therapy compared to the baseline, while gaining 0.2 kg after four months of treatment. Moreover, patients with positive sputum status did not gain weight during the first two months of therapy. According to Yohannes et al., (2013) study in Gondar, Ethiopia, Multivariate and bivariate logistic regression analysis was conducted to evaluate the significance of association between PTB and explanatory variables. Socio demographic variables such as sex; residence and occupational status of the respondents were not significantly associated with pulmonary tuberculosis infection in this study.

Another study done by Hiwot A.et al. (2013), in Dessie, bivariate logistic regression was used to identify possible explanatory (independent) variables and those variables, which have a p-value of less than 0.05, were taken to logistic regression. As a result, age (P = 0.011), presence of TB patients in the family (P = 0.012), educational status (P = 0.019), dose (P=0.0001),marital status (P = 0.021), smoking (P = 0.010), duration of diabetes (P = 0.008),and consumption of alcohol (P = 0.002) were significantly associated with development of PTB. On the other hand, place of residence (P = 0.141), religion (P = 0.649), monthly income (P = 0.666), HIVstatus (P = 0.920), sex (P = 0.103), blood glucose level (P = 0.267), and occupational status (P = 0.659) were significantly associated with the occurrence of pulmonary TB. The result obtained from Worodria et al (2011) and C-S Wang. (2008) the older age was the risk factor of pulmonary tuberculosis and loss of body weight of TB patients.

According to Xuefeng Liu and Michael J. Daniels (2003) study body weight and smoking status were inversely associated each other. The results showed that Age, moderate-intensive exercise, dose and presence of TB patients in the family were associated with the longitudinal out comes.

## 2.2. TB Treatment Regimen

Despite the fact that designing retreatment regimens for patients with TB and a history of category I treatment is a cornerstone in TB management, few studies have addressed this issue. Currently, WHO recommendations are based on category II regimen for retreatment of these cases. However, the successful outcome of this regimen is relatively low; according to a study in Morocco showed that the mean retreatment success rates of the category II regimen were, 58.0% and 51.4% respectively, among failure and default cases (Ottmani et al., 2006, Tabarsi et al., 2008). The prevalence of MDR TB in patients with CAT I failure or a history of more than one course of an irregular category I anti-TB regimen, which were 56% and 55%, respectively. Therefore, it is evident that introducing treatment regardless of drug susceptibility test (DST) pattern may be an improper approach to patients, especially those who failed or had irregular category I treatment (Tabarsi et al 2008).

A retreatment strategy based on DST and replacing the category II regimen may improve clinical outcomes among category I treatment failures, a great part of who are patients with MDR TB.The strategy significantly reduces delays in arriving at MDR TB diagnosis and the initiation of MDR TB therapy (Tabarsi et al., 2008). The management of children with TB should be in line with the Stop TB Strategy, taking into consideration the particular epidemiology and clinical presentation of TB in children(WHO, 2006) Obtaining good treatment outcomes depends on the application of standardized treatment regimens according to the relevant diagnostic category, with support for the child and carer that maximizes adherence to treatment. A recent development in treatment recommendations is that, following a comprehensive literature review, ethambutol is now considered safe in children at a dose of 20 mg/kg (range 15–25 mg/kg) daily (WHO,2006).

Adverse events caused by anti-tuberculosis drugs are much less common in children than in adults (WHO 2006), in addition to TB treatment regimen dosing dedicated for infants and children is more accurate because its calculation is based on body weight, so all these factors justify better treatment outcome in this group of patients (WHO 2006). The outcome of the standard retreatment regimen for TB is poor, particularly in those infected with both HIV and MDR-TB.This indicates that standard retreatment approach to TB as implemented in low and

middle income setting with high prevalence HIV is inadequate and stresses the importance of new, more effective strategy (Jones Lopez et al., 2011*)*.

## 2.3. Overview of Models for Longitudinal Outcomes

### 2.3.1. Linear Mixed Model

Pinheiro, J. C. and Bates, D. M. (2000 ) used Linear mixed model a special case of continuous repeated data, characterized as having between-subject and within-subject variation, time dependent covariates and missing data. This can accommodate these complex features of longitudinal data whereas traditional methods are limited by statistical assumptions. More importantly, the approach allows for explicit modeling of the variation between subjects and within subjects. The term "mixed-effects" refers to the expression of the model into fixed effects and random-effects. The linear mixed-effects model assumes that the observations follow a linear regression where some of the regression parameters are fixed or the same for all subjects, while other parameters are random, or specific to each subject (Laird and Ware, 1982).

### 2.3.2. Generalized Linear Mixed Model

Linear mixed models (which incorporate random effects) and generalized linear model is also used to handle normal data, but it is more general than the LMM which assume normality. It is also used for non normal data by using link functions and exponential family [e.g. normal, Poisson or binomial] distributions). GLMMs are the best tool for analyzing no normal data that involve random effects: all one has to do, in principle, is specify a distribution, link function and structure of the random effects. The inclusion of random effects in the linear predictor reflects the idea that there is natural heterogeneity across subjects or clusters in some of their regression coefficients (Antonio & Beirlant, 2006).

According to McCulloch clarification, GLMM is very versatile in that they can handle non-normal data, nonlinear models, and a random effects covariance structure. This can be used to incorporate correlations in models, model the correlation structure, identify sensitive subjects and can be used to handle heterogeneous variances. The modeling process is relatively straightforward, requiring the following decisions: what is the distribution of the data, what is to be modeled, what are the factors, and are the factors fixed or random? This

all makes GLMM attractive for use in modeling. Unfortunately, computing methods for much of the class of GLMM is an area of active research.

### 2.3.3. Joint Modeling Approach

Longitudinal studies typically involve following one or more cohorts of subjects or experimental units repeatedly over two or more time points. Multivariate longitudinal studies are comprised of repeated responses each of which consists of two or more elements. In a multivariate longitudinal model, there are two types of correlations. One, called serial correlation, is between observations at different time points within a subject and the other, called cross correlation, is between observations on different response variables at each time point. If different types of outcomes are measured at each time point, the correlation structure is more complicated and hence, more difficult for drawing inference. Separate analyses of the different types of outcomes can lead to biased inferences because of those correlations. Therefore, it is more desirable to jointly model multivariate outcome variables of different types together. As many studies measure multiple response outcomes of different types for each subject repeatedly, there are many approaches to model the different outcomes jointly (Olkin, I. and R. F. Tate ,1961; Zeger and Liang, 1986).There are two general approaches for modeling multivariate longitudinal observations with differing outcome types. One proposed method for formulating the joint distribution of different types of outcomes is to model the relationship between the different outcomes using random effects. In this approach, different mixed models for each outcome are joined by imposing a common distribution for their random effects. It allows their model-specific random effects to be correlated, and this model allows for flexible correlation patterns. This model has a disadvantage of the high-dimensionality of the vector of random effects as the number of outcome variables gets large.

Another approach is using the product of the marginal distribution of one of the responses and the conditional distribution of the remaining response given the other response, that is,

$$f(y_{continuous}, y_{discrete}) = f(y_{continuous)} f(y_{continuous} | y_{discrete}) = f(y_{discrete)} f(y_{discrete} | y_{continuous})$$

Here, $f(.)$denotes the probability density functions associated with the outcomes. In the conditional model, one has to choose an outcome to condition on which plays the role of a time-varying covariate. Thus, two possible types of models can lead to very different results

depending on whether the conditioning variable is a discrete or a continuous outcome. The main disadvantages with conditional modeling approach are that it is hard to get easy expressions for the association between both continuous and discrete outcomes, and that it does not directly lead to marginal inference. Also, if we have more than two outcomes, there will be many more possible factorizations instead of only the two associated with two outcomes. Hence, a conditional model is often not the preferred choice for an analysis of high-dimensional multivariate longitudinal data (Gueorguieva, R., 2013).

Catalano and Ryan (1992) described a joint distribution for bivariate clustered binary and continuous outcomes by factorizing the marginal distribution of a continuous outcome and a conditional distribution of a binary outcome given the continuous outcome. They used the concept of a latent variable. The type of latent variable used by Catalano and Ryan supposed that an unobserved continuous variable underlies the observed binary variable. Hence, they assume that a binary outcome results from dichotomizing the continuous latent variable. Accordingly, they used a linear link function for the marginal distribution of the continuous outcome and used a correlated probit model for the conditional distribution of the binary outcome.

Gueorguieva and Agresti (2001) used an approach similar to Catalano and Ryan (1992) for joint model. They studied a correlated probit model that applies an underlying latent normal variable for the binary outcomes but use a random effects model instead of a conditional model. The focus of their work was on the joint, subject-specific effects on the models. Tsiatis, DeGruttola, and Wulfsohn (1995) examined the relationship between the CD4 count and survival time in patients with acquired immune deficiency syndrome (AIDS). They proposed a two-stage procedure by plugging the estimates from longitudinal models into a Cox proportional hazards model. Another study conducted through correlated random effect is the works of Regan.et al. (1999) and Gueorguieva, R.V., and Sanacora, G (2006) focused on malformation and fetal weight which are typical primary endpoints for live offspring. Given a sub sampling of fetal weight, to show how valid estimates could be obtained when the two longitudinal outcomes are correlated. Possible association between weight changes during treatment and treatment outcome has been investigated in some studies Hoa, N.B.et al. (2012).

# CHAPTER-THREE

# METHODS

## 3.1. Study Population and Design

The data were extracted from the retrospective cohort follow up chart of TB patients' between September (2011) to July (2013) from Jimma University Specialized Hospital. This chart was recorded by assigning an identification number per individual and contains epidemiological and clinical information of all tuberculosis patients. The data consists of four hundred five individuals, measured repeatedly at least two times on each patient. Five data collectors (extractors), one supervisor for five days were allocated. Training was given for both supervisor and data collectors on how the data were coded and recorded.

## 3.2. Data Source and Description

In this thesis, secondary data were obtained from four hundred five individuals those who were included in the study to evaluate the weight variation over time and it association with sputum status those who were under patients' follow-up records. Patients registered for treatment still six months in Jimma University Specialized Hospital on category of TB, Sex, Age, HIV status and Residence have been taken.

The anti-tuberculosis regimens used for Category I and III patients were 2ERHZ/6RH and for Category II (Re-treatment regimen) patients was 2S (ERHZ) / 1(ERHZ) / 5E3 (RH), (H = isoniazid; R = rifampicin; Z = pyrazinamide; E = ethambutol; S = streptomycin. Numbers before the acronyms indicate the duration of the treatment phase in months and numbers in subscript (for Category II only) indicate the number of times the drug is given each week whereas the drug is given daily for Category I and III patients). Treatment for category I and II patients was extended by another month if the sputum smear remained positive at the end of Intensive Phase.

## 3.3. Variables of Interest

The variable of interest that was considered in the analysis are the response (dependent) and the explanatory (independent) variables.

### 3.3.1. Dependent Variable

The following two response variables were considered to be studied simultaneously:

- Bodyweight recorded in kilogram (kg) from treatment start (baseline) and repeatedly measured in a two months basis. And
- Sputum status of TB patients which is dichotomized as 1 if the sputum status is positive and as 0 if negative (1 =Positive, 0 = Negative)

### 3.3.2. Predictor Variables (Independent variable)

The predictor variables also called covariates. These covariates are categorical and continuous. The predictor (covariate) variables which were assumed to influence the weight and sputum status of TB patients included in the model are:

- ✓ Sex
- ✓ Age
- ✓ Types of tuberculosis (pulmonary and pulmonary smear negative tuberculosis)
- ✓ HIV status
- ✓ Dose
- ✓ Residence
- ✓ Time

Table 1: Description of Predictor Variables included in the Analysis

| variables | Representation | Coding |
|---|---|---|
| Sex | $X_1$ | 1=Male, 0=Female |
| Age | $X_2$ | continuous |
| Types of tuberculosis | $X_3$ | 1=Pulmonary, 0= pulmonary smear neg. |
| HIV status | $X_4$ | 0= HIV Negative,1= HIV Positive |
| Dose | $X_5$ | Continuous |
| Residence | $X_6$ | 0=rural,1=semi-urban, urban=2 |
| Time | $X_7$ | Continuous |

## 3.4. Exploratory Data Analysis

The first step in analyzing longitudinal data is to explore the data given. Observe patterns through graphical displays and summary statistics that are relevant to the research question. Diggle et al.(2002) recommends illustrating relevant raw data as much as possible, identifying both cross-sectional and longitudinal patterns that may be of interest, and identifying outliers or unusual observations. Variability trends within subjects and between subjects will help in choosing a covariance structure for the model as explained in the next section.

- Individual profiles
- Mean structure
- Variance function
- Correlation structure

### 3.4.1. Individual Profile Plot

A natural way to explore longitudinal profiles is by plotting individual profiles. This is extremely helpful as additional tool in the selection of appropriate models.

### 3.4.2. The Average Evolution

The average evolution describes how the profile for a number of relevant subpopulations (or the population as a whole) evolves over time. The results of this exploration will be useful in order to choose a fixed-effects structure for the linear mixed model.

### 3.4.3. The Variance Structure

In addition to the average evolution, the evolution of the variance is important to build an appropriate longitudinal model. Clearly, one has to correct the measurements for the fixed-effects structure and hence raw residuals must be used. The variance function must be relatively stable and hence a constant variance model could be a plausible starting point.

### 3.4.4. The Correlation Structure

The correlation structure describes how measurements within a subject correlate. The correlation function depends on a pair of times and only under the assumption of stationary does this pair of times simplify to the time lag only. This is important since many exploratory and modeling tools are based on this assumption. If one or both structures are varying with time, the standardized residuals will contribute useful additional information. A different way of displaying the correlation structure is using a scatter plot matrix.

## 3.5. Statistical Models

### 3.5.1. Models for a Single Longitudinal Continuous Response

#### 3.5.1.1. Linear Mixed Model

The linear mixed model (LMM) and its corresponding marginal models are appropriate statistical models for continuous data, given that they duly acknowledge dependence between observations within subjects, through the use of random effects. In this case, body weight of TB patients is a continuous response of interest for linear mixed model.

Let the random variable $\mathbf{Y_{ij}}$ denote the continuous response of interest (body weight), for the $i^{th}$ patient, measured at the $j^{th}$ time point. $i = 1,2, \dots\dots\dots, N, j = 1,2, \dots\dots\dots, n_i$.

Where, $Y_i$ be a p-dimensional vector of all repeated measurements for the $i^{th}$ TB patient that is $\boldsymbol{Y_i} = (\boldsymbol{Y_{i1}}, \boldsymbol{Y_{i2}} \dots\dots\dots\dots, \boldsymbol{Y_{ini}})^t$. The LMM is specified as:

$$Y_{ij} = X_{ij}{}^t\beta + Z_{ij}{}^t b_i + \varepsilon_{ij} \qquad (1)$$

This model involves two set of covariates $X_{ij}$ and $Z_{ij}$. The $n_i \times p$ covariates $X_{ij}$ are associated with a p-dimensional vector of fixed-effects parameters $\beta$ and the $n_i \times q$ set of covariates $Z_{ij}$ associated with the random effects $b_i \sim N(0, D)$.

In addition, $\varepsilon_{ij} \sim N(0, R_i)$ represents the residual of the $i^{th}$ patient at time j. Given the random effects $b_i$, the residuals $\varepsilon_{ij}$ are often (but not always) assumed independent. The variance-covariance matrix D indicates the degree of heterogeneity of subjects. When all dependent residuals are considered, a variety of covariance structures are then possible for both D and $R_i$, such as unstructured, compound symmetry, and first-order autoregressive matrices. Note that: $E(y_{ij}) = E\left(E(y_{ij}|b_i)\right) = X_{ij}^t\beta$ and so marginal and conditional parameters are equal. Alternatively, one can postulate the following marginal model:

$$Y_{ij} = X_{ij}{}^t\beta + \varepsilon_{ij}^* \qquad (2)$$

$$\varepsilon_{ij}^* \sim N(0, V_i^*).$$

The marginal distribution of the response is then, $Y_{ij} = N\left(X_{ij}{}^t\beta, V_i^*\right)$. In this case, correlation is taken into account through covariance parameters in$V_i^*$. Again, different specifications of the covariance structure can be imposed for the covariance $V_i^*$ as mentioned above for D and$R_i$. It is well known that the marginal model resulting from (1) is a special case of (2). Linear mixed model therefore implies a specific marginal model with the hierarchical linear mixed model therefore implies a specific marginal model with $\varepsilon_{ij}^* \sim N(0, V_i^*)$. Where,$V_i = z_i D z_i^t + R_i$.A very important fact is that the implied marginal model removes the positive definiteness restrictions on the D and $R_i$ matrices, merely requiring that $V_i$ be positive definite.

#### 3.5.1.1.1. Covariance Structure of Linear Mixed Model

In longitudinal analysis, the most important thing is the selection of the covariance matrix. In fact, choosing an appropriate covariance structure is the first step in model selection (Hedeker & Gibbons, 2006). When choosing a covariance structure, all covariates of interest should be

included in the model since the significance tests of the covariates depend on the covariance structure (Hedeker & Gibbons, 2006). The covariates in the model are to remain the same through the testing of different covariance structures for a proper comparison. Testing can be done using the AIC criterion. Some common variance-covariance matrices include (Hedeker & Gibbons, 2006; Vonesh & Chinchilli, 1997) are give below:

- **Variance components (VC):-** The VC structure is the standard variance components and is default structure.

$$\begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & 0 \\ 0 & 0 & 0 & \sigma_4^2 \end{bmatrix}$$

- **Autoregressive (1):-** The AR (1) structure has homogeneous variances and correlations that decline exponentially with distance. It also means that two measurements that are right to next to each other in time are going to be pretty correlated (depending on the value of $\rho$), but that as measurements get farther and farther apart they are less correlated.

$$\sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

- **Compound symmetry (CS):-** The CS structure is well-known compound symmetry structure required for split plot designs "in the old days". In CS structure the variances are homogeneous. There is a correlation between two separate measurements, but it is assumed that the correlation is constant regardless of how far apart the measurements are.

$$\begin{bmatrix} \sigma^2 + \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma^2 + \sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma^2 + \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma^2 + \sigma_1^2 \end{bmatrix}$$

- **Unstructured (UN):-** The UN structured is the most "liberal" of all allowing every term to be different. It requires fitting the most parameters of any structure, t (t+1)/2.

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 & \sigma_{13}^2 & \sigma_{14}^2 \\ \sigma_{12}^2 & \sigma_2^2 & \sigma_{23}^2 & \sigma_{24}^2 \\ \sigma_{13}^2 & \sigma_{23}^2 & \sigma_3^2 & \sigma_{34}^2 \\ \sigma_{14}^2 & \sigma_{24}^2 & \sigma_{34}^2 & \sigma_4^2 \end{bmatrix}$$

- **TOEPLITZ:-**The TOEP structure is similar to the AR(1) in that all measurements next to each other have the same correlation, measurements two apart have the same correlation different from the first, measurements three apart have the same correlation different from the first two, etc. However, the correlations do not necessarily have the same pattern as in the AR (1). Technically, the AR (1) is special case of the Toeplitz.

$$\begin{bmatrix} \sigma^2 & \sigma_1^2 & \sigma_2^2 & \sigma_3^2 \\ \sigma_1^2 & \sigma^2 & \sigma_1^2 & \sigma_2^2 \\ \sigma_2^2 & \sigma_1^2 & \sigma^2 & \sigma_1^2 \\ \sigma_3^2 & \sigma_2^2 & \sigma_1^2 & \sigma^2 \end{bmatrix}$$

Heterogeneous versions of the above are a simple extension. That is the variances, along the diagonal of the matrix, do not have to be the same. Note that this adds more parameters to be estimated, one for every measurement.

### 3.5.2. Generalized Linear Model

Generalized linear model (GLM) is a flexible generalization of ordinary least squares regression. In linear mixed model, there are several assumptions such as normality, homoscedasticity and linearity. However, in GLM we will give up such kind of assumptions: drop the normality in favor of the exponential family of distributions; abandon the homoscedasticity in favor of a known function, which is called variance function, and explain how the individual variation depend on the respective mean; throw away the linearity assumption in favor of a known function which is called the link function and then translate the nonlinearity into a function of linear relationships which we call linear predictors.

### 3.5.2.1. Models for a Single Longitudinal Binary Response

#### 3.5.2.1.1. The Generalized Linear Mixed Model (GLMM)

The generalized linear mixed model is the most frequently used random effects model in the context of binary repeated measurements. Not only is it a rather straightforward extension of the generalized linear model for univariate data to the context of clustered measurements, there is also a wide range of software tools available for fitting these models. In this study, estimation and inference for this class of random-effects models will be seen in particular.

Suppose that $Y_{ij}$ is an outcome for the $i^{th}$ patient, measured at the $j^{th}$time point, and $b_i$are assumed to be normally distributed with mean 0 and variance-covariance matrix D, that is $b_i$ ~N(0,D), with $E(b_{i)}$= 0 and Var($b_i$) = D. Then, it is assumed that the conditional distribution of the response $Y_{ij}|b_i$ is independent and belongs to the following exponential family density

$$f_i(y_{ij}/b_i,\phi) = exp\{\phi^{-1}[y_{ij} - \psi(\theta_i)] + c(y,\phi)\} \quad (3)$$

The expected value of $E(Y_{ij}/bi) = \boldsymbol{\mu_{ij}}$ and there is a link function g (.) that relates the conditional mean of the data to the linear predictor$gE(Y_{ij}/bi) = \eta_{ij}$.

**Model formulation**

Let $Y_{ij}$ is categorical response variable sputum status follows a binomial distribution i.e. $\boldsymbol{Y_{ij}}\sim\boldsymbol{Bin}(\boldsymbol{n_j},\boldsymbol{\mu_{ij}})$ that belongs to the exponential family with the density function of the form (3). The logit or logistic function is

$$g\left(\mu_{ij}\right) = X_{ij}^t\beta + Z_{ij}^t b_i \ \ or \qquad logit\left(\mu_{ij}\right) = X_{ij}^t\beta + Z_{ij}^t b_i \quad (4)$$

Where, $\mu_{ij}$: The mean of $Y_{ij}$ which is related to the covariates of X by link function

$X_{ij}$: Covariates of the i$^{th}$ patient of the j$^{th}$ time point

$\beta$: Regression coefficients of$X_{ij}$.

$Z_{ij}$: The covariates of the random effects of the i[th] patient at j[th] time

$b_i$:The random effect which are assumed to be multivariate normal distribution having mean vector 0 and covariance matrix G, i.e. $b_i \sim N(0, G)$

In the GLMM, $\beta_j$ is the increase in log-odds of negative result for any patient associated with a one-unit increase in$X_{ij}$. Averaging across individuals $\beta_j$ is also the increment in log-odds for the population, because the mean value of $b_i$ in the population is zero. But, $\exp(\beta_j)$ is not the average multiplicative effect on the odds in the population, because the average of an antilog is not the same as the antilog of an average. The nonlinearity of the link function requires us to make a distinction between the meaning of $\beta_j$ in the marginal and multilevel analyses. It is said to be subject-specific. A very lucid and thorough discussion on the differences between population-average and subject-specific effects is given by Fitzmaurice et al. (2004).

### 3.5.3. Joint Model for Continuous and Binary Responses

Joint modeling is a term used to reflect a modeling approach whereby two response processes are linked via a common set of random effects. It can be used to model two related outcomes such as a count and a binomial variable, two count outcomes, or two binomial outcomes, both of which have some correlated effect; or to model a survival and recurrent event process; or, to model a survival and longitudinal variable. Under the joint modeling framework, we may, for example, use one process to inform the second, with the main emphasis being on analysis of one of the processes; alternatively, we may be interested in analyzing both outcomes jointly and using the correlated random effect structure to better inform both processes. Basically, the broad objective of joint modeling is to provide a framework for analyzing the systematic relationship among multiple outcomes while appropriately accounting for the correlation among these outcomes. The association among the two outcomes is captured by correlating the normal random effects describing the continuous and binary outcome, respectively. Several authors have developed joint models for analysis of multivariate longitudinal data using latent normal variables (Daniels and

Normand, 2006; Dunson, 2003; Gueorguieva and Sanacora, 2006). In this section, the joint model for the association of weight and sputum status has been modeled as follow.

**Joint model formulation**

Let $Y_{1ij}$ longitudinal continuous outcome (body weight) at the j[th] time point on the i[th] subject and $Y_{2ik}$ longitudinal binary outcome (sputum status) at $k^{th}$ time point on the $i^{th}$ subject, with densities, $f_{1i}(y_{ij})$ and $f_{2i}(y_{ik})$ respectively$i = 1,2 \dots \dots \dots, N, j = 1,2, \dots \dots, n_{1i}$ $and\ k = 1,2, \dots \dots, n_{2i}$.Formulation of a joint model could be based on the random-effects approach for $Y_{1ij}$ and $Y_{2ik}$ are modeled separately by including subject-specific random-effects $b_{1i}$and $b_{2i}$ respectively. Conditionally upon the random-effects, the two outcomes are assumed independent. Hence, the association between $Y_{1ij}$ and $Y_{2ik}$ is captured by letting $b_{1i}$and $b_{2i}$to be correlated (Molenberghs and Verbeke, 2005). A special case is the so-called shared or correlated parameter model, where the same set of random-effects is assumed for all outcomes. However, this approach has the disadvantage that it is based on strong assumptions about the association of the two outcomes, and hence may not be valid (Molenberghs and Verbeke, 2005).The joint model elements from the linear mixed model of Sections 3.5.1.1. and the generalized linear mixed model of section 3.5.2.1.1, in one single model, the so called joint model, conditional upon the random effects, has the following

$$f_i(y_{1i}, y_{2i}|b_{1i}, b_{2i}, \beta, \alpha, \mu_{2ik}) = f_{1i}(y_{1i}|b_{1i}, \beta) \times f_{2i}(y_{2i}|b_{2i}, \alpha, \mu_{2ik}) \quad (5)$$

The model entities are defined as follows:

$Y_{1ij}$=the body weight for the i[th] subject at time j (continuous).

$Y_{2ik}$=the sputum status for the i[th] patient at time k (binary).

$$Y_{1ij} = X_{1i}^t(t)\beta + Z_{1i}^t(t)b_{1i} + \epsilon_{1ij}$$

$$Y_{2ik}|b_{2i} \sim Bernoulli(Pr(y_{2ik} = 1))$$

$$logit(\Pr(y_{2ik} = 1)) = X_{2i}^t(t)\alpha + Z_{2i}^t(t)b_{2i}$$

This model for the complete observed data $Y_{2ij}$ and $Y_{2ik}$ translates into the following model for condition on the $b_{1i}\ and\ b_{2i}$

$$\begin{cases} \mu_{1ij} = X_{1ij}^t\beta + Z_{1ij}^t b_{1i} \\ \eta(\mu_{2ik}) = X_{2ik}^t\alpha + Z_{1ik}^t b_{2i} \end{cases} \quad (6)$$

Where, $\mu_{1ij}$ and $\mu_{2ik}$ are the conditional mean of for the two observed variables and $g(.)$ is known link function.

The random effects and random errors are assumed to be normally distributed.

$$b_i = (b_{1i}, b_{2i})^t \sim N[0, D]$$

$$\epsilon_i = (\epsilon_{1i}, \epsilon_{2i})^t \sim N(0, \Sigma_i) = N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_{11}^2 & \sigma_{12}\\ \sigma_{12} & \sigma_{22}^2\end{pmatrix}\right]$$

Where, D, the covariance matrix of the random effects, has the following structure:

$$\begin{pmatrix}b_{10}\\b_{11}\\b_{20}\\b_{21}\end{pmatrix} \sim MVN\left[\begin{pmatrix}0\\0\\0\\0\end{pmatrix}, \Sigma = \begin{pmatrix}\sigma_{d10}^2 & \rho_{10,11}\sigma_{d10}\,\sigma_{d11} & \rho_{1020}\sigma_{d10}\,\sigma_{d20} & \rho_{1021}\sigma_{d10}\,\sigma_{d21}\\ & \sigma_{d11}^2 & \rho_{1120}\sigma_{d11}\,\sigma_{d20} & \rho_{1121}\sigma_{d11}\,\sigma_{d21}\\ & & \sigma_{d20}^2 & \rho_{2021}\sigma_{d21}\,\sigma_{d20}\\ & & & \sigma_{d21}^2\end{pmatrix}\right]$$

**Association of the two outcomes**

One important question that may be addressed with a joint model is how the evolution of one response is associated with the evolution of another response ("association of the evolutions"). The correlation between the evolutions for the two random effects is given by: i≠j.

$$r_{AOE} = \frac{cov(b_{1i}, b_{2j})}{\sqrt{var(b_{1i})}\sqrt{var(b_{2j})}} \quad (7)$$

### 3.5.4. Parameter Estimation Methods

Parameter estimation is always one of the most important aspects of statistical inference for any model. Many techniques have been made for parameter estimation for linear mixed model (LMM), Generalized Linear mixed Model (GLMM) and Joint model.

#### 3.5.4.1. Parameter Estimation of LMM

Estimation for separate mixed effect model**:** - Estimation of the parameters in LMM is usually based on maximum likelihood (ML) or restricted maximum likelihood (REML) estimation for the marginal distribution of $Y_i$ which can easily be seen to be $Y_i \sim N(X_i\beta, Z_i DZ_i^{'} + \Sigma_i)$. Note that model LMM implies a model with very specific mean and covariance structures, which may or may not be valid, and hence needs to be checked for every specific data set at hand. Note also that, when $\Sigma_i = \sigma^2 I_{ni}$ , with $I_{ni}$ equal to the identity matrix of dimension $n_i$, the observations of subject i are independent conditionally on the random effect $b_i$. The model is therefore called the conditional independence model. Even in this simple case, the assumed random-effects structure still imposes a marginal correlation structure for the outcomes $Y_{ij}$. Indeed, even if all $\Sigma_i$ equal $\sigma^2 I_{ni}$ , the covariance matrix in $Y_i \sim N(X_i\beta, Z_i DZ_i^{'} + R)$ is not a diagonal matrix, illustrating that, marginally, the repeated measurements $Y_{ij}$ of subject $i$ are not assumed to be uncorrelated. The marginal mean (expected value) and marginal variance-covariance matrix of the vector $Y_i$ is equal to: $E(Y_i) = X\beta$ and $Var(Y_i) = V_i = Z_i DZ_i^{'} + R$

##### 3.5.4.1.1. Maximum Likelihood Estimation

Suppose a random sample of N observations is obtained from a linear mixed effect model as defined above, and then the likelihood of the model parameters, given the vector of N observations, is defined as:

$$L = l(\beta, \theta, Y_i) = \prod \left\{ 2\pi^{\frac{-n_i}{2}} \det(V_i)^{\frac{-1}{2}} \exp\left(\frac{-1}{2}(Y_i - X_i\beta)' V_i^{-1}(Y_i - X_i\beta)\right)\right\}$$

Then, the MLE of $\widehat{\beta}$ on combining all the information from all the N subjects equals.

$$\hat{\beta} = \left(\sum X_i V_i^{-1} X_i\right)^{-1} \left(\sum X_i V_i^{-1} Y_i\right)$$

Where, det refers to the determinant and the elements of the $\boldsymbol{V}_i$ matrix are functions of the covariance parameters in Θ.

#### 3.5.4.1.2. Restricted Maximum Likelihood Estimation

The REML estimation method applies ML estimation techniques to the likelihood function. The only difference is the REML estimation method is associated with a set of "error contrasts" rather than associated with the original observations. Therefore, it will lose degrees of freedom and give less biased estimates of the variance components. The bias issue cannot be neglected, especially when the number of parameters is not small relative to the total number of observations. Let us start the easier case. We consider the estimation of $\sigma^2$ for the general linear model. The MLE of $\hat{\sigma}^2$ is $\hat{\sigma}^2 = \frac{(Y-X)^t(Y-X\beta)}{n}$. Where, $\hat{\beta} = (X^t X)^{-1} X^t Y$. The REML estimate of $\hat{\sigma}^2$ is the minimum variance unbiased estimator $\hat{\sigma}^2 = \frac{(Y-X)^t(Y-X\beta)}{n-p}$.

REML: $L(D,\Sigma) = -\frac{1}{2} log|V| - \frac{1}{2} log|X^t V^{-1} X| - \frac{1}{2} m V^{-1} m - \frac{n-p}{2} \log(2\pi)$

Where, $m = Y - X(X^t V^{-1} X)^{-1} X^t V^{-1} Y$ and p is the rank of X Estimating fixed effect (β ) and random effect (b ) parameters in the Mixed Model. Once getting estimates of D and ∑, which are denoted by, $\hat{D}$ and $\hat{\Sigma}$ respectively, and D is nonsingular, we solve mixed model equations.

$$\begin{pmatrix} X^t\hat{\Sigma}^{-1}X & X^t\hat{\Sigma}^{-1}Z \\ Z^t\hat{\Sigma}^{-1}X & Z^t\hat{\Sigma}^{-1}Z + \hat{D}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} X^t\hat{\Sigma}^{-1}Y \\ Z^t\hat{\Sigma}^{-1}Y \end{pmatrix}$$

Then the solution is

$$\hat{\beta} = \left(X^{-1}\hat{V}^{-1}X\right)^{-1} X^{-1}\hat{V}^{-1}Y$$

$$\hat{b} = \hat{D}Z^t V^{-1}(\hat{Y} - X\hat{\beta})$$

If $\hat{D}$ is singular, then the mixed model equations are modified (Henderson 1984) as follows

$$\begin{pmatrix} X^t\hat{\Sigma}^{-1}X & X^t\hat{\Sigma}^{-1}Z \\ \hat{L}Z^t\hat{\Sigma}^{-1}X\hat{L} & Z^t\hat{\Sigma}^{-1}\hat{L}+I \end{pmatrix}\begin{pmatrix} \hat{\beta} \\ \hat{\tau} \end{pmatrix}=\begin{pmatrix} X^t\hat{\Sigma}^{-1}Y \\ \hat{L}Z^t\hat{\Sigma}^{-1}Y \end{pmatrix}$$

Where $\hat{L}$ is the lower triangle cholesky root of$\hat{D}$, satisfying$\hat{D}=\hat{L}\hat{L}^t$.Both $\hat{\tau}$ and a generalized inverse of the left hand side coefficient matrix are then transformed using $\hat{L}$ to determine $\hat{b}$ from this$var(\hat{\beta})=(X^t\hat{V}^{-1}X)^{-1}$.

### 3.5.4.2. Parameter Estimation of GLMM

**Gaussian Quadrature: -** The Gaussian Quadrature approximates the integral of a function, with respect to a given kernel, by a weighted sum over predefined abscissas for the random effects. Unlike other numerical integration techniques, the abscissas are spaced unevenly throughout the interval of integration. With a modest number of Quadrature points, along with appropriate centering and scaling of the abscissas, the Gaussian Quadrature approximation can be highly effective (Abramowitz and Stegun 1964). In the particular context of random-effects models, so-called adaptive Quadrature rules can be used (Pinheiro and Bates, 2000), where the numerical integration is centered on the estimates of the random effects, and the number of Quadrature points is then selected in terms of the desired accuracy. To illustrate the main ideas, we consider Gaussian and adaptive Gaussian Quadrature, designed for the approximation of integrals of the form $\int f(z)\phi(z)dz$ for a known function f (z) and for $\phi(z)$ the density of the multivariate standard normal distribution. Therefore first standardize the random effects such that they get the identity covariance matrix.

The likelihood contribution of subject i is

$$f_i\left(y_{ij}|\beta,bi,\phi\right)=\int\prod_{j=1}^{ni}f_{ij}\left(y_{ij}|b_i,\beta,\phi\right)f(b_i|D)db_i \qquad (8)$$

From this, the likelihood for β, D and $\phi$ is given as

$$L\left(\beta,D,\phi\right)=\prod_{i-1}^{N}\int\prod_{j=1}^{ni}f_i\left(y_{ij}|b_i,\beta,\phi\right)f(b_i|D)db_i \qquad (9)$$

### 3.5.4.3. Parameter Estimation of the Joint Model

Parameters in the joint model are estimated using maximum likelihood, based on

$$L(\beta, \alpha, D) = \prod_{i=1}^{N} f_i\,(Y_{1i} = y_i, Y_{2i} = 1) =$$

$$f_i(y_{ij}, y_{ik}) | b_{1i}, b_{2i}\beta, \alpha, \mu_{2ik} = \frac{1}{(2\pi)^{\frac{n_i}{2}} |\Sigma_i|^{\frac{1}{2}}} \chi e^{\frac{-1}{2}\left[(Y_{1i} - X_{1i}^t \beta - Z_{1i}^t b_{1i})^t \Sigma^{-1} (Y_{1i} - X_{1i}^t \beta - Z_{1i}^t b_{1i})\right]}$$

$$\chi \prod f(y_{ik} | b_{2i}, \alpha, \mu_{2ik})$$

Even though this analytical joint marginal likelihood can be maximized, it is cumbersome to manipulate. It is therefore more convenient to maximize the likelihood after employing numerical techniques, rather than to integrate out the random-effects distribution. Gaussian and adaptive Gaussian Quadrature are designed for such purpose, up to a pre-specified level of accuracy (Pinheiro and Bates 1995, 2000). The standard errors of the parameter estimates are computed from the inverse Hessian matrix (second derivatives) at the estimates obtained numerically. Major statistical tools, such as the SAS procedure NLMIXED, are readily available for fitting the models specified in this paper.

### 3.5.5. Model Comparison Technique

The primary objective of model comparison is to choose the saturate model that provides the best fit to the data. In order to select the best and final model which is appropriately fits with the given longitudinal data, it is necessary to compare the different models by using different techniques and methods. Hence, models are compared with Akaki Information Criteria (AIC), the Bayesian Information Criteria (BIC), and the Likelihood ratio test methods for nested were used at 5% level of significance. Both Linear mixed models and joint were compared using AIC, BIC and Likelihood ratio test and GLMM models were compared using AIC and Likelihood ratio test.

**Akaike's information criterion (AIC)** is a measure of goodness of fit of an estimated statistical model. It is not a test on the model in the sense of hypothesis testing; rather it is a tool for model selection. The AIC penalizes the likelihood by the number of covariance parameters in the model, therefore

AIC= -2Log (L) +2p

Where, L is the maximized value likelihood function for the estimated model and p is the number of parameters in the model. The model with the lowest AIC value is preferred

**BIC=-2log Likelihood + nP log (N)**

Where, -2 logL is twice the negative log-likelihood value for the model

P: - is the number of estimated parameters.

N: - is the total number of observations used to fit the model. Smaller values of AIC and BIC reflect an overall better fit

**Likelihood ratio test:** it is constructed by comparing the maximized log likelihoods for the full and reduced models respectively and the test statistic is

$$T_{LR}{}^2 = -2ln\lambda_N = -2ln\left(\frac{L_{ML}(\hat{\alpha}_{ml,0})}{L_{ML}(\hat{\alpha}_{ml})}\right)$$

Where, $\hat{\alpha}_{ml,0}$ and $\hat{\alpha}_{ml}$ are respective maximum likelihood estimates which maximize the likelihood functions of the reduced and full model. The asymptotic null distribution of the LR test statistic is a chi-square distribution with degrees of freedom equal to the difference between the numbers of parameters in the two models.

### 3.5.6. Model Diagnosis

After a mixed effects models have been fitted it is important to check whether the underlying distributional assumptions for the random effects and the residuals appear valid for the data. Diagnostic methods for linear models are well established. The most useful method for diagnostics, according to Pinheiro and Bates (2000), are based on plots of the residuals, the fitted values and Normal Q-Q plot of estimated random effects is an important method for checking the normality (Myers et al., 2010). In this thesis, all diagnostic methods can be used by using the functions qqnorm.lme and plot.lme in Pinheiro et al. (2010). Here the standardized, or Pearson residuals, defined as the raw residuals divided by the estimated corresponding standard deviation, are used.

## 3.6. Ethical Consideration

Ethical clearance for the study was obtained from the Ethical Clearance Review Board of College of Natural Science, Jimma University. In addition, official letter of co-operation was written to Jimma University Specialized Hospital to get permission for accessing the data.

# CHAPTER -FOUR

# ANALYSIS AND RESULTS

## 4.1. Baseline Information

The data consists of 405 patients who were under Tuberculosis treatment between 2011 and 2013 in Jimma University Specialized Hospital. Two response variables were considered; continuous longitudinal outcome body weight and binary longitudinal outcome sputum status which were measured approximately every two months; approximately equal number of patients were visited at each follow up time. During TB diagnosis time 35.5% and 64.5% of male and female patients were visited respectively. The numbers of male patients were relatively higher than that of females. HIV negative patients account for the highest proportion (83.5%) and 55.5% of TB patients were suffered from pulmonary tuberculosis and 44.5% were suffered from pulmonary smear negative tuberculosis. The urban group showed the highest percentage (46.9%) with respect to the frequency of visits than the other two categories. The mean of baseline weight is approximately 51k.g with standard deviation 11.102k.g. and at the end of follow up time are 53.1 and 11.1 respectively. The mean and standard deviation of the weight of patients corresponding to the given covariates is summarized in Table 2.

The proportion of tuberculosis patients those were visited during the follow up as positive and negative sputum status is 39.3% and 60.7% respectively. The percentage of patients whose sputum conversion is positive at base line is 13.2% and the percentage of patients whose sputum conversion is positive at the end of follow up is 1.3. Similarly, the proportion of tuberculosis patients whose sputum status is negative at base line and the end of follow up time are 11.8% and 23.6% respectively.

Table 2.The mean and standard deviation of body weight over time according to the corresponding covariates

| | Sex | | Residence | | | Types of TB | | HIVS | |
|---|---|---|---|---|---|---|---|---|---|
| | female | male | rural | smurban | urban | PTB | PNTB | pos | neg |
| **Mean** | 47.6 | 53.7 | 52.7 | 51.6 | 51.0 | 51.3 | 52.2 | 50.2 | 51.8 |
| **St.dev.** | 11.41 | 10.5 | 11.8 | 11.0 | 11.1 | 12. 5 | 9.8 | 12.4 | 11.0 |

Table3. Proportion of Sputum Conversion with baseline Categorical Covariates

| | | Sputum status | |
|---|---|---|---|
| Variable | Categories | Positive (100%) | Negative (100%) |
| Sex | male | 181(10.8) | 869(51.7) |
| | female | 96(5.7) | 534(31.8) |
| Residence | rural | 82(4.9) | 322(19.2) |
| | Semi-urban | 96(5.7) | 430(25.6) |
| | urban | 140(8.3) | 640(39.3) |
| Types of TB | Pulmonary TB | 234(13.9) | 576(34.3) |
| | P.Smear negative | 27(1.6) | 843(50.2) |
| HIV status | positive | 207(12.3) | 235(14) |
| | negative | 239(14.2) | 1000(59.5) |

From Table 3, it is clear to see that a very large proportion of male patients had negative sputum status than female patients. The percentage of negative sputum status among patients whose residence is rural is smaller than those patients whose residence is semi-urban and those whose residence is urban. Also, the percentage of negative sputum status among HIV positive patients is the smaller while it is the larger among HIV negative patients. Finally,

from the table it is possible to say that the sputum status being positive for Pulmonary Smear negative tuberculosis patient's decreases.

## 4.2. Separate Analysis of Continuous Longitudinal Outcome (Weight)

### 4.2.1. Exploratory Analysis

#### 4.2.1.1. Individual profiles plot of Body Weight

Individual profile plots of weight over time have been explored to identify general trends within and between subjects and may detect change over time that provides information about the variability at given time (in figure 4.1) and the evolution over time has been observed.

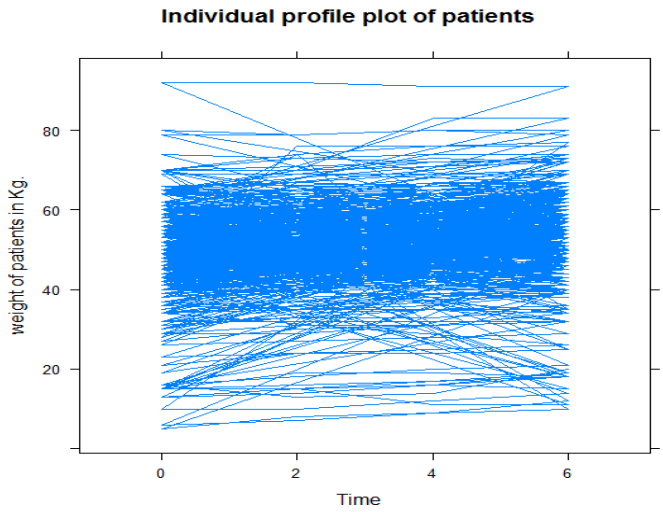



Figure 4.1.Individual profile plot of Body weight of TB patients

As illustrated in figure 4.1, the plots indicate that variability of the weight of tuberculosis patients is somewhat the same at baseline and at the end of follow up time. But, the profiles plot shows that there is between and within variability of patients' weight which implies that the between and within subject differences must be considered.

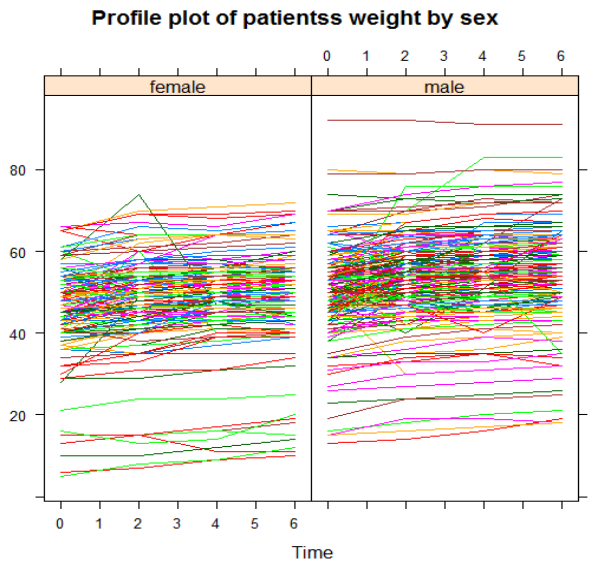



Figure 4.2.Individual plot of weight by Sex

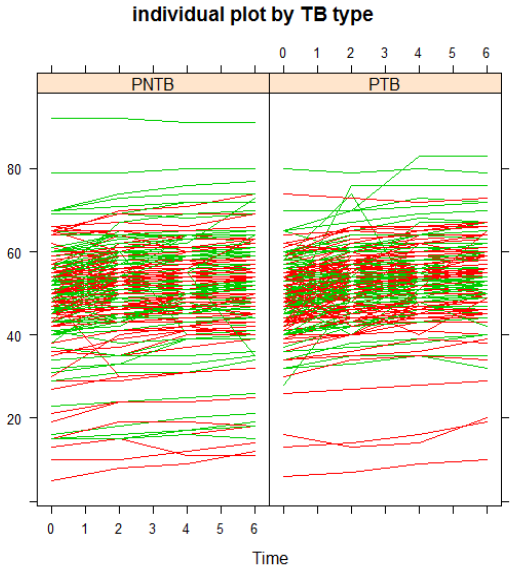



Figure 4.3.Individual plot of Body weight of CTB

In figure 4.2, it seems that the weight variation for males' is higher than females' .The weight of most patients' is increasing over time for both subjects. And from figure 4.3, the weight variation for pulmonary smear negative patients' is higher than pulmonary tuberculosis patients.

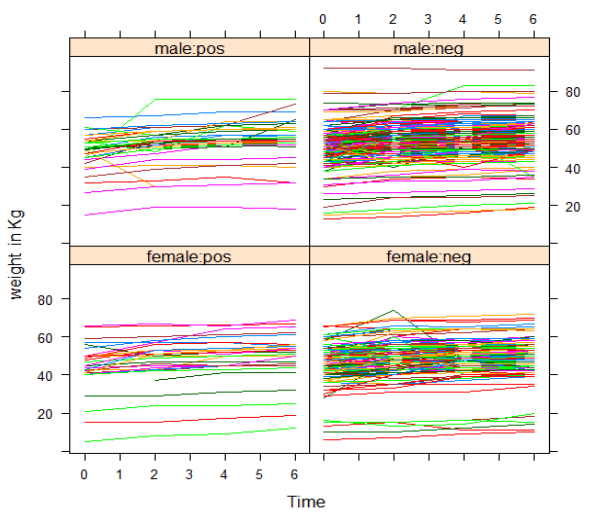



Figure 4.4.Individual plot by Sex: HIVstatus

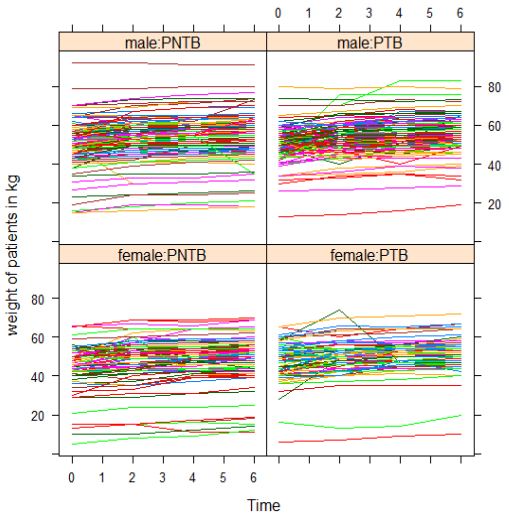
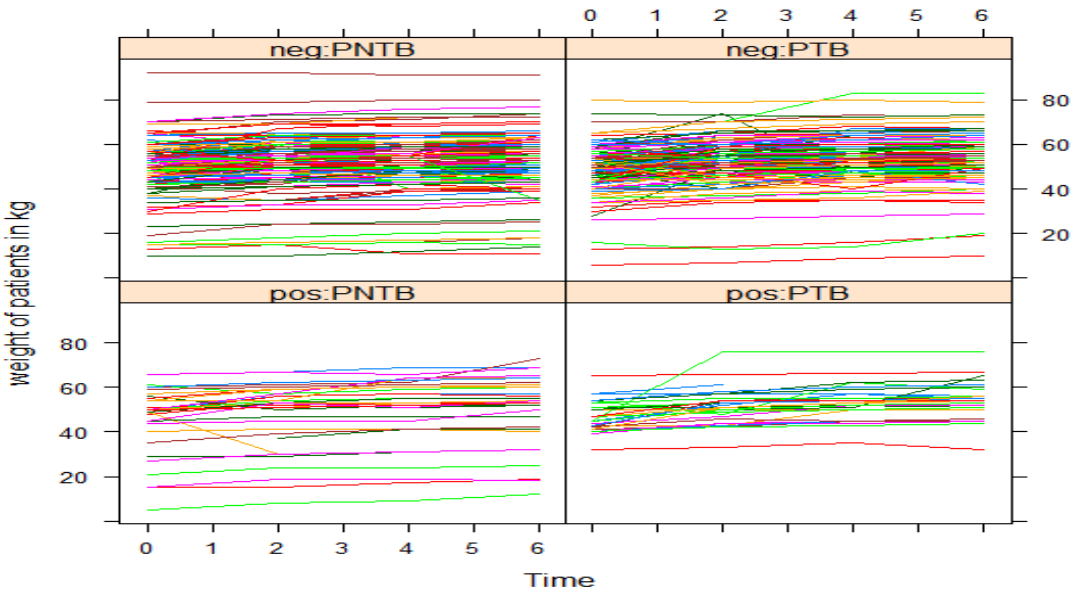



Figure 4.5.Individual plot of weight by sex: HIVstatus

In figure 4.4, there is the same variability both sexes having a negative and a positive HIV status, but a higher variability is shown in figure 4.5 among females that have suffered from pulmonary tuberculosis.

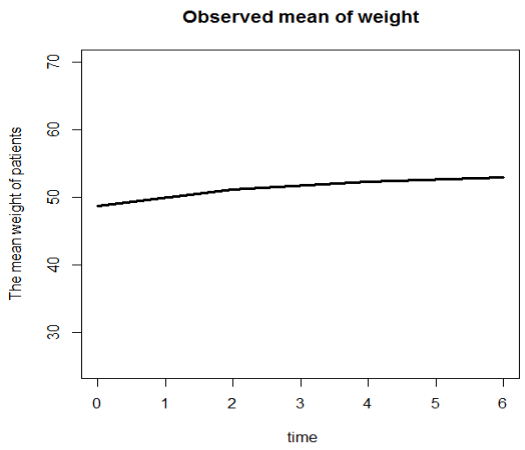



Figer 4.6.Individual plot of the weight of TB patient by HIV status and Types of TB

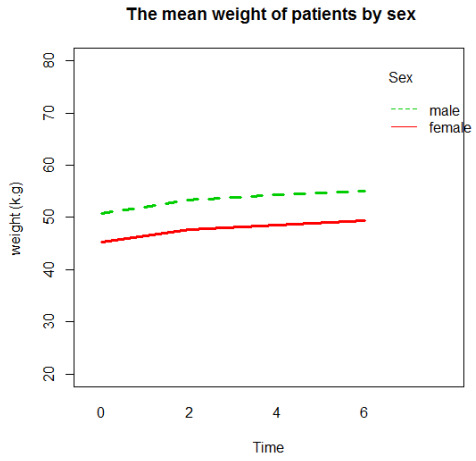
From figure 4.6, higher variability is shown among HIV positive patients those suffered from pulmonary smear negative tuberculosis.

#### 4.2.1.2. Exploring Mean Structure of Body Weight of Patients'

The average evolution describes how the weight of patients evolves over time. It is used to build an appropriate longitudinal model for weight of TB patients**.**

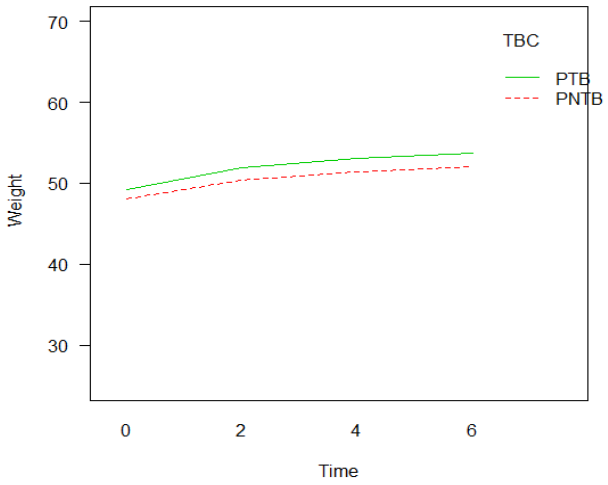



Figer 4.7.The mean profile plot over time




Figer 4.8.The mean profile plot by sex

Figure 4.7 shows the mean profile plot of TB patients, the mean weight of patients' seems like to a linear relationship with time. Hence, linear effects of time may be important. As shown in figure 4.8, the mean weight of male patients' is higher than females. It appears that there is a linear relation with time



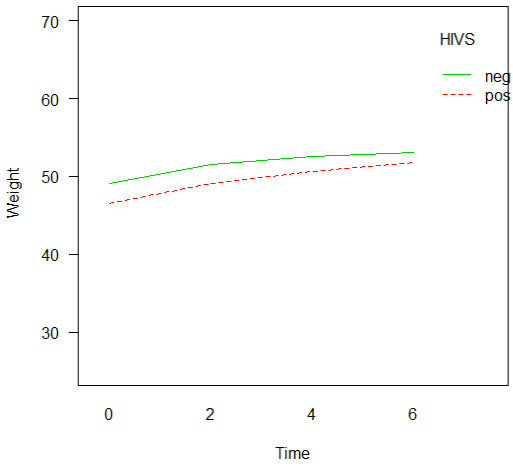
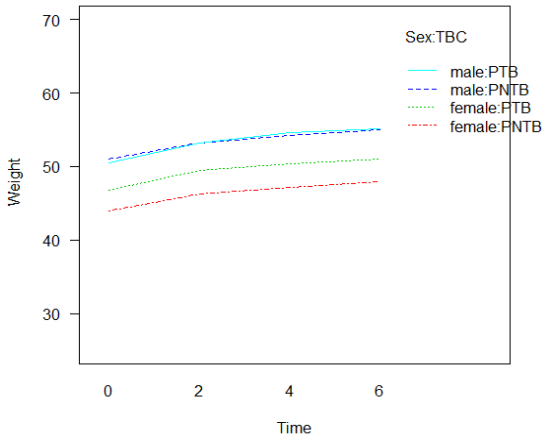




Figure 4.9.The mean profile plot by types of TB Figure 4.10.The mean profile plot by HIV status

The mean weight of pulmonary tuberculosis patients is higher than the mean weight of pulmonary smear negative tuberculosis patients as shown figer 4.9. And also from figure 4.10 the mean weight of HIV negative patients' higher than those patients whose HIV status is positive. The mean of weight has linear relation over time.



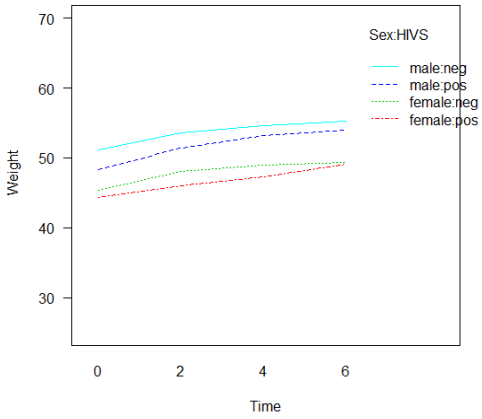




Figure 4.11.The mean profile plot by Sex: TB Figure 4.12.The mean profile plot by Sex: HIV status

From the above figure 4.12, mean profile plot of weight by the interaction of HIV status and sex, the mean weight of male patients' those who are HIV negative is higher than those males whose HIV status is positive. Similarly, the mean weight of female patients' those who are HIV negative is higher than those females whose HIV status is positive. Also, figure 4.11 shows the mean profile plot of the weight of TB patient by the interaction of categories of TB and sex. As a result, the mean weight of male pulmonary tuberculosis patients is higher than the mean weight of male pulmonary smear negative patients'. In the same way, the mean weight of female pulmonary tuberculosis patients' is higher than the mean weight of female pulmonary smear negative patients'.

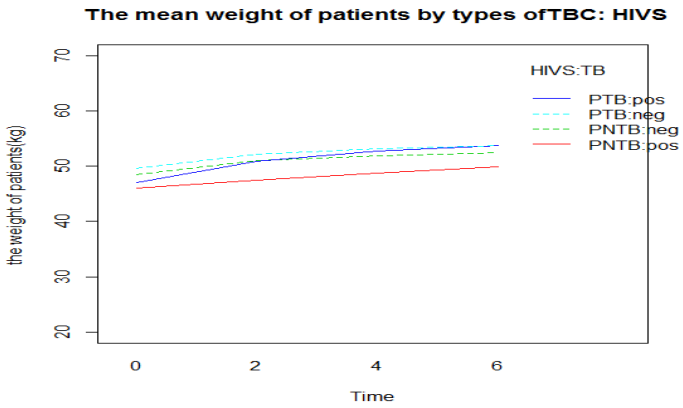



Figure 4.13.The mean profile plot by category of TB and HIV status

The mean weight of HIV negative pulmonary tuberculosis patients is higher as depicted in figure 4.13.

### 4.2.1.3. Exploring the Variability of Weight of TB patients'

In addition to the mean evolution, the variability also used to build appropriate longitudinal data. Hence, the variability plot is given in figure 4.14 having different categories.

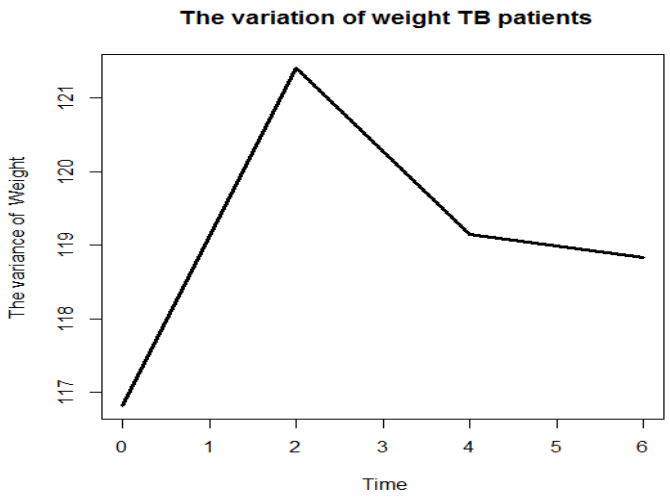



Figure4.14.The variance profile plot of the weight of TB patient over time

As shown form figure 4.14, the variability of weight of patients is not constant. It is increased until the second month and decreased after the second month.

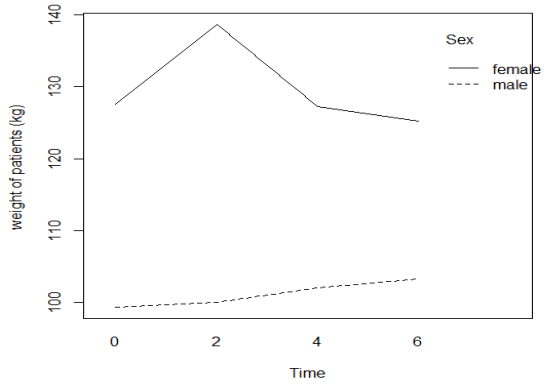



Figure4.15. The variance plot by sex

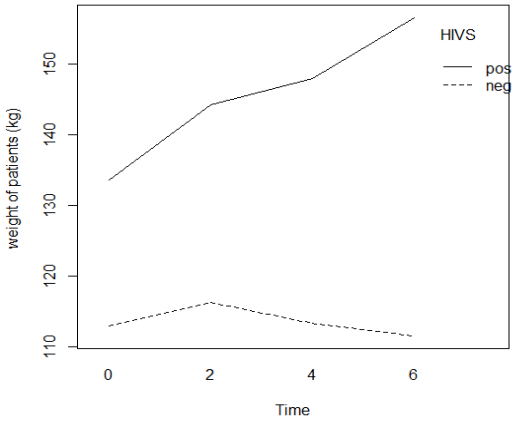
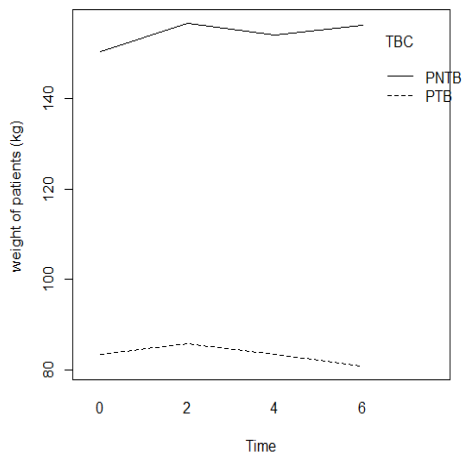
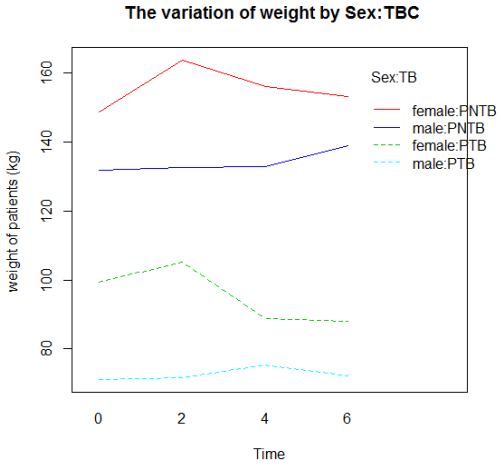



Figure4.16. The variance plot by HIV status

The variance of male patients is higher than females from figure 4.15 and the variability of HIV positive patients' is higher than the variability of HIV negative tuberculosis patients (Figure4.16).




Figer 4.17.The variance by TB category




Figer 4.18.The variance by Sex: TB category

In figure 4.17, the variability of pulmonary smear negative is higher than the variability of pulmonary tuberculosis patients. In such a way, the variability of female pulmonary smear negative tuberculosis patients is higher and the variability of male patients those suffered from pulmonary tuberculosis is lower from figure 4.18.

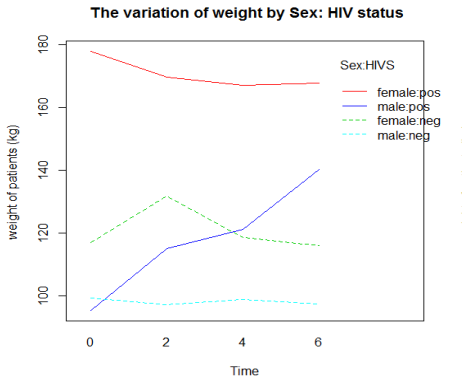



Figure 4.19.The variance plot by sex and HIV status

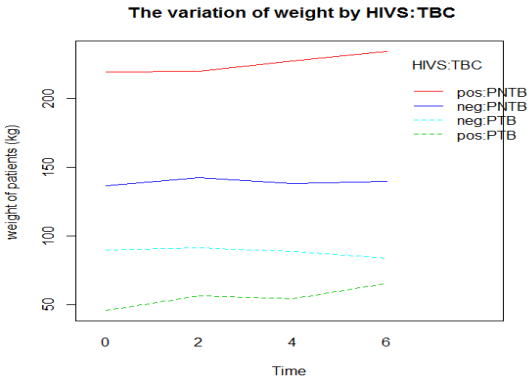



Figure 4.20.The variance by HIV status and TB category

From figure 4.19, the variance of HIV positive female patients is high and male whose HIV status is negative is low. The variability of HIV positive pulmonary smear negative tuberculosis patients is higher in figure4.20.

#### 4.2.1.4. Exploring the Correlation Structure

The first step in the model building process for a linear mixed-effects model, after the functional form of the model has been decided, is choosing which parameters in the model, if any, should have a random-effect component included to account for between-group variation. The Lmlist function and the methods associated with it are useful for this. From the individual profiles and mean structure of these data linear relationship of weights as a function of time seems suitable. In order to make the plot visible, the interval plot of the subset data has been depicted below whose ID number is less 26274.

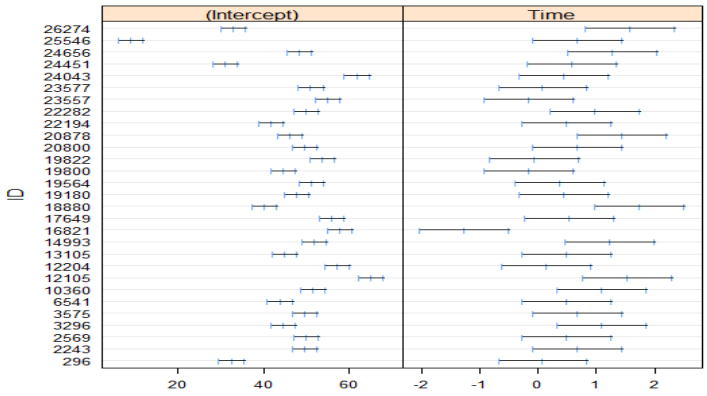



Figure 4.21: Interval Plots for Subject Specific Intercept and Slope of body Weight

As depicted in figure 4.21, the lines are not overlapping each other. Hence, both the random intercept and slopes are important to fit an appropriate linear mixed model for the body weight of tuberculosis patients'. The correlation matrix and the scatter plot of the weight of TB patients shown below:

Table 4: The correlation matrix of the weight of TB patients

| | **W1** | **W2** | **W3** | **W4** |
|---|---|---|---|---|
| **W1** | 1.0000000 | 0.9394148 | 0.8514068 | 0.8418190 |
| **W2** | 0.9394148 | 1.0000000 | 0.8856854 | 0.8765999 |
| **W3** | 0.8514068 | 0.8856854 | 1.0000000 | 0.9851858 |
| **W4** | 0.8418190 | 0.8765999 | 0.9851858 | 1.000000000 |

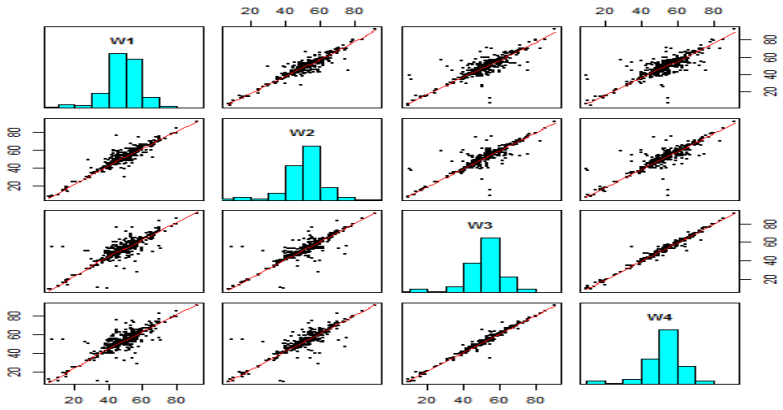



Figure 4.22: Scatter Plot Matrix for the Weight of TB patients

Pair wise scatter plots were used for exploring the correlation between any two repeated measurements of patients' weight and it appears that there is a positive relationship between patients' weight taken at different time as shown in figure 4.22.

### 4.2.2. Separate Linear Mixed Model for Body Weight

Linear mixed model is appropriate model for repeated continuous longitudinal data and the appropriate model has been model as follow:

#### 4.2.2.1. Random Effect Selection

As shown figure 4.22, the exploring of random effect, both random intercept and slope are important. To select the random effect to the model, intercept only, slope only and both intercept and slope different models would be fitted and compared (table5). An appropriate random effect to the model must be selected by using the model selection criterion like AIC, BIC and likelihood ratio test. The small p-values corresponding to the fitted model indicates that the model is preferable. Similarly, with small AIC and BIC value is considered as the best model.

Table5: Selection of random effect to be included in LMM for body weight

| | Model | df | AIC | BIC | logLik | Test | L.Ratio | p-value |
|---|---|---|---|---|---|---|---|---|
| model.int | 1 | 26 | 11280.92 | 11426.78 | -5614.461 | | | |
| Model.slope | 2 | 28 | 12402.73 | 12559.80 | -6173.363 | 1 vs 2 | 1117.804 | <.0001 |
| Model.both | 3 | 31 | 11040.72 | 11214.62 | -5489.357 | 2 vs 3 | 1368.011 | <.0001 |

As shown table 5, three models have been fitted. These were with random intercept only (model.int), random slope only (Model.slope) and both random intercept and slope (Model.both). The result is similar as illustrated in figure 4.22, because, both Akaki information criteria (AIC) and Bayesian information criteria (BIC) values of the model fitted from both random intercept and slope are smaller than the other models, which implies that both random intercept and slope are important.

Table6. Comparison of linear time effect model with quadratic time effect

| Effect | model | df | AIC | BIC | logLik | Test | L.Ratio | p-value |
|---|---|---|---|---|---|---|---|---|
| Linear | 1 | 20 | 8578.390 | 8922.37 | -5401.02 | | -506.52 | |
| Quadratic | 2 | 31 | 9101.618 | 9285.05 | -4281.46 | 1vs 2 | -450.132 | <.0001 |

From table5, both random intercept and slope were taken to account in linear mixed model. Having these random effects, the comparison of linear and quadratic time effect has done in table6. The smaller AIC value is the better the model. The AIC value of a linear time effect model is smaller than a quadratic time effect (8578.390<9101.618).Hence, linear time effect is important in favor of quadratic time effect. This is supported the mean evolution of body weight in figure 4.1, Therefore, time has linear random effect.

#### 4.2.2.2. Linear Mixed Model for Body Weight with Linear Time Effect

To select the fixed effect for body weight of TB patients', for all covariates and interaction terms were fitted. The large p-values relative to the level of significance ($\alpha$=0.05) corresponding to the given covariates showed the covariates are not significant. Eliminating all

insignificant covariates and interaction terms step by step was the right way to handle the model with all significant variables which contains the smallest AIC and BIC values. The insignificance covariate eliminated were residence (semi-urban (p= 0.76065), urban (p= 0.94416) and time*residence semi-urban (p= 0.68305) and time*residence urban (p= 0.85618). The maximum likelihood estimates of the parameter estimates are summarized in table 7 below. Thus, Linear mixed model with linear time effect is fitted as follow using the most important covariates which have been selected by removing insignificant covariates step by step and comparing the models based on AIC values:

$$Weight_{ij} = \beta_0 + \beta_1 Sex_i + \beta_2 \text{Age}_i + \beta_3 \text{TBC}_i + \beta_4 \text{HIVS}_i + \beta_5 \text{dose}_i + \beta_6 \text{Time}_{ij} + \beta_7 Sex_i : \text{Time}_{ij} + \beta_8 \text{Age}_i : \text{Time}_{ij} + \beta_9 \text{TBC}_i : \text{Time}_{ij} + \beta_{10} \text{HIVS}_i : \text{Time}_{ij} + \beta_{11} \text{dose}_i : \text{Time}_{ij} + b_{0i} + b_{1i} \text{Time}_{ij} + \varepsilon_{ijk}$$

Where,

$Weight_{ij}$= body weight on the $i^{th}$ patients at the $j^{th}$ time point

$\text{Time}_{ij}$ = time in which the body weigh of the $i^{th}$ patient obtained

$\text{Resd}_i$ = the residence in which of the $i^{th}$ patient live(rural, urban and semiurban)

$\text{Sex}_i$ = sex of the $i^{th}$ patient (male, female )

$\text{HIVS}_i$ = the HIV status of the $i^{th}$ patient live(negative, positive)

$\text{Age}_i$ = the age of the $i^{th}$ patient

$\text{TBC}_i$ = TB categories in which the $i^{th}$ patient suffered from (pulmonary tuberclosis, pulmonary smear negative)

$\text{dose}_i$ = the amount of dose in which the $i^{th}$ patient obtained

$\beta_0$ = the mean weight at baseline

$\beta_1, \ldots\ldots\ldots, \beta_{12}$ Coefficients of Fixed Effect

$b_{0i}$ = random intercept and $b_{1i}$ Coefficients of random Effect time

$\varepsilon_{ijk}$ = $random\ Error\ Term$

.

Table7. Parameter estimates and standard errors for the separate LMM for the body weight of the final model using ML and REML

| | ML(Std.errors) | REML(St.err) | 95%CI | p-value |
|---|---|---|---|---|
| Intercept | 29.689(1.452) | 29.691 (1.4518) | (28.1436 31.016) | 0.0000 |
| Sexmale | 6.166(0.714) | 6.156 (0.7143) | (2.6339 8.11612) | 0.0000 |
| Age | 0.187(0.006) | 0.1869 (0.0028) | (0.1138 0.26038) | 0.0000 |
| TBCPTB | 0.130(0.0743) | 0.1304 (0.07425) | (0.0146 1.9266) | 0.0073 |
| HIVSneg | 0.796(0.0774) | 0.7951 (0.7737) | (0.1191 1.4845) | 0.0057 |
| dose | 3.727(0.432) | 3.7268 (0.43165) | (2.7931 4.8317) | 0.000 |
| Time | 1.782(0.498) | 1.60922 (0.4927) | (0.1663 1.7406) | 0.0022 |
| Sexmale: Time | -0.1807(0.198) | -0.1806 (0.1975) | (-0.2981 -0.0223 ) | 0.0085 |
| Age: Time | 0.015(0.0487) | 0.0148 (0.0487) | (0.0116 0.0195) | 0.0090 |
| TBCPTB:Time | -0.216(0.186) | -0.2168(0.1855) | (-0.2246 -0.0576) | 0.0015 |
| HIVSneg:Time | -0.305(0.257) | -0.3053 (0.257) | (-0.4331 -0.0499) | 0.0045 |
| dose: Time | -0.252(0.137) | -0.2517 (0.1369) | (-0.1457 -0.0498) | 0.0022 |

As shown table7 above, the estimates and standard errors of the two methods are almost the same; this is due to that the large sample size and small number of parameters of the final model.

#### 4.2.2.3. Pattern of Variance-Covariance Structure

Similarly with the mean pattern over time, correlation structure also used to select the best model. Among different correlation structures, in this thesis, unstructured covariance model, compound symmetric covariance models, and autoregressive structure of order one, AR (1), Toeplitz, Variance component were used and compared.

Table8. Comparison of model with different correlation function for weight

| Covariance structure | -2LL | AIC | BIC |
|---|---|---|---|
| Compound Symmetry (CS) | 11365.2 | 9357.4 | 9369.9 |
| **Unstructured(UN)** | **10351.7** | **9340.2** | **9363.1** |
| Autoregressive (AR(1)) | 10371.3 | 9387.2 | 9390.1 |
| Variance component(VC) | 12354.6 | 9452.7 | 9489.2 |
| Toeplitz (TOEP) | 13924.2 | 9553.2 | 96930.2 |

As it is shown in table 8, among different covariance structure mentioned, the model with unstructured covariance structure was preferred for the continuous outcome weight with respective small values of AIC and BIC of 9340.2, and 9363. Other variance structures didn't give an improvement for the model over the fitted model. The assumption of normality for the within-group errors was assessed with the normal probability plot of the residuals, produced by the qqnorm method and was satisfied(Appendix-1) .

The mean weight of tuberculosis at base line is approximately 30 kg. Sex is significantly associated with body weight of TB patients; the average baseline weight of males is higher by nearly 6 kg than that of females (p-value=0.0000). Age of the patients is also significantly and positively associated with weight as one year increments of the age of patients the mean weight also increased by 0.187kg((p <0.0001)) .The mean weight of patients those who are HIV negative have 0.796 kg higher than those who are HIV positive patients (p=0.0057). A one-tab increase in the dose, the mean change of the weight of patients at base line is 3.727 kg. As the follow up time increases, the mean change of weight is 1.5kg (p=0.000). Similarly, the mean weight of tuberculosis patients those who have suffered from pulmonary tuberculosis is 0.130kg higher than those patients suffered from pulmonary smear negative tuberculosis by adjusting the other covariates(p=0.007). Like the interpretation the main effects of covariates the interaction terms have been done as follow: Since the coefficient of sex: time interaction is -0.181, Over time, the rate of change of weight among males is lower by 0.18 kg as compared to females after adjusting other covariates. In the same way, the rate of change of weight of pulmonary tuberculosis is lower by 0.216kg as compared to pulmonary smear negative tuberculosis and the rate change of weight of among HIV negative patients is lower by 0.305kg compared from HIV positive.

**Assumption of Random effect**

Table9. Standard errors and covariance structure for random effects (LMM)

| Effects | parameters | Standard deviation |
|---|---|---|
| Var ($b_{10}$) | $d_{00}$ | 8.284 |
| Var ($b_{11}$) | $d_{11}$ | 0.527 |
| Cov($b_{10}$, $b_{11}$) | $d_{01=}$ $d_{10}$ | -0.121 |
| Var($\varepsilon_{ij}$) | $\sigma_1^2$ | 2.177 |

Now, the variance of random slope for the linear time effect $d_{11}$=0.527 shows a small variability among the linear time effect compared from the variability of intercept effect. The assumption of random effects are normally distributed, with mean zero and covariance matrix Σ and are independent for different groups. The two basic diagnostic plots qqnorm normal and pairs scatter were used to investigate the validity of assumption two. Basically, qqnorm normal plot of estimated random effects was used for checking marginal normality and identifying. Plots of random intercept versus slope did not suggest any departures from the assumption of homogeneity of the random effects distribution as drawn in (Apendix-1).

## 4.3. Separate Analysis of Binary Longitudinal Outcome (Sputum Conversion)

### 4.3.1. Generalized Linear Mixed Model for Sputum Conversion

Under the GLMM, model fitting began by adoption of the marginal model covariates. Additionally, the model also included the random effects in this case, random intercepts and slope to address the between and within- variations. First, all main effect covariates and the random intercepts and slope model were fitted and as usual, non significant covariates were removed sequentially starting from variables with highest p-value for fixed effect covariates. Having $\mu_{2ik}$ denoted the probability of positive sputum status for $i^{th}$ patient at the $j^{th}$ time point then, the saturated models for GLMM has been fitted as follows:

$$logit(\mu_{2ik}) = \alpha_0 + \alpha_1 \text{Age}_i + \alpha_2 \text{TBC}_i + \alpha_3 \text{dose}_i + \alpha_4 \text{Time}_{ij} + \alpha_5 \text{Age}_i : \text{Time}_{ij} + \alpha_6 \text{TBC}_i : \text{Time}_{ij} + \alpha_7 \text{dose}_i : \text{Time}_{ij} + b_{2i}$$

Where,

$\text{Time}_{ij}$ = time in which the body weigh of the $i^{th}$ patient obtained

$\text{Age}_i$ = the age of the $i^{th}$ patient

$\text{TBC}_i$ = TB categories in which the $i^{th}$ patient suffered from

(pulmonary tuberclosis, pulmonary smear negative)

$\text{dose}_i$ = the amount of dose in which of the $i^{th}$ patient obtained

$b_{2i}$ = random effect,Which assumed to be normal distributed having mean vector 0 and covariance matrix G $i.e.\ b_{2i} \sim N(0, G)$

Table 10: Parameter estimates and standard errors for GLMM

| Effects | Estimate(S.e) | p-value | 95%CI |
|---|---|---|---|
| Intercept | 1.2097(0.06460) | 0.0494 | (0.1164 2.5358) |
| Age | 0.0137 (0.2526) | 0.0001 | (0.0011 0.0264) |
| CTBPTB | -2.604(0.1865) | 0.0058 | (-3.3390 -1.869) |
| dose | -0.1704 (0.7464) | 0.0067 | (-0.3359 -0.0049) |
| Time | -2.022(0.0131) | 0.0080 | (-3.4850 -0.5590) |
| Age*Time | 0.0018 (0.3112) | 0.0001 | (0.0007 0.0290) |
| Time*CTBPTB | 1.3271(0.2112) | 0.0002 | (0.7172 1.9370) |
| dose*Time | 0.0188 (0.1362) | 0.0032 | (0.0048 0.0328 ) |
| $\sigma_2^2$ | 1. 657(0.0646) | 0.0049 | (1.1919 2.1221) |

As shown table 10, all covariates have a significant effect on a positive sputum status of tuberculosis patients. The estimate for age is 0.0137 with standard error of 0.2526.The pulmonary tuberculosis estimate and standard errors are 2.6040 and 0.1865 respectively. The

between-patient variance of positive sputum status is estimated to be 1.657 which is significantly different from zero (p=0.0049) at 5% implied heterogeneity has been ignored.

### 4.4. Joint Model of Weight and Sputum Conversion

The separate model was fitted for TB data as introduced in section 4.2.2 and 4.3 where body weight as well as sputum status were measured repeatedly for each patients. The two outcomes were modeled jointly to capture association between them. For the continuous longitudinal outcome, the covariate of base line age, sex, categories of TB, dose, time and HIV status were included. For the binary longitudinal outcome, the same covariates were included except sex and HIV status. Using appropriate model selection techniques, the saturated model has been obtained as follow from linear mixed effect and generalized linear mixed effect model.

Denote by $Y_{1ij}$ and $Y_{2ik}$ weight and sputum status on the i$^{th}$ patient at the j$^{th}$ and k$^{th}$ time point, respectively. And the mean of binary outcome sputum status is $\mu_{2ik}$ .The LMM and GLMM model have been formulated for these data as:

$$\begin{aligned} Y_{1ij} = {} & \beta_0 + \beta_1 Sex_i + \beta_2 \mathrm{Age_i} + \beta_3 \mathrm{TBC_i} + \beta_4 \mathrm{dose_i} + \beta_5 \mathrm{HIVS_i} + \beta_6 \mathrm{Time_{ij}} + \beta_7 Sex_i : \mathrm{Time_{ij}} \\ & + \beta_8 \mathrm{Age_i} : \mathrm{Time_{ij}} + \beta_9 \mathrm{TBC_i} : \mathrm{Time_{ij}} + \beta_{10} \mathrm{HIVS_i} : \mathrm{Time_{ij}} + \beta_{11} \mathrm{dose_i} : \mathrm{Time_{ij}} \\ & + \mathrm{b_{10i}} + \mathrm{b_{11i}} \mathrm{Time_{ij}} + \end{aligned}$$

$$\begin{aligned} logit(\mu_{2ik}) = {} & \alpha_0 + \alpha_1 \mathrm{Age_i} + \alpha_2 \mathrm{TBC_i} + \alpha_3 \mathrm{dose_i} + \alpha_4 \mathrm{Time_{ij}} + \alpha_5 \mathrm{Age_i} : \mathrm{Time_{ij}} \\ & + \alpha_6 \mathrm{TBC_i} : \mathrm{Time_{ij}} + \alpha_7 \mathrm{dose_i} : \mathrm{Time_{ij}} + \mathrm{b_{20i}} + \mathrm{b_{21i}} \mathrm{Time_{ij}} \end{aligned}$$

Further, assume that the subject-specific random effects of continuous and binary longitudinal outcomes $\mathrm{b_{1i}}$ and $\mathrm{b_{2i}}$ are multinormally distributed as:

$$\begin{pmatrix} b_{1i} \\ b_{2i} \end{pmatrix} \sim MVN \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right]$$

The result of maximum likelihood estimation of the fitted saturated model has been summarized below

Table11**:** A parameter estimates and standard errors of the joint model for body weight and sputum conversion

| | Weight | | Sputum conversion | | |
|---|---|---|---|---|---|
| Effects | Estimates (S.e) | p-value | Effects | Estimate(S.e) | p-value |
| (Intercept) | 29.978(1.4514) | 0.0000 | intercept | 1.2718 (0.07269) | 0.0043 |
| Sexmale | 6.276(0.7041) | 0.0004 | Age | 0.03027 (0.0297) | 0.0075 |
| Age | 0.188 (0.003) | 0.0166 | CTBPTB | -2.8508 (0.0255) | 0.0001 |
| CTBPTB | 0.142 (0.0958) | 0.0004 | dose | -0.1907(0.0431) | 0.0154 |
| HIVSneg | 0.798 (0.0655) | 0.0001 | Time | -2.553 (0.0420) | 0.0018 |
| dose | 3.816(0.2703) | 0.0141 | Time: Age | 0.096(0.2107) | 0.0058 |
| Time | 1.886(0.04133) | 0.0001 | Time: CTB | 2.4947(0.0217) | 0.0001 |
| Sexmale: Time | -0.179(0.0776) | 0.0049 | dose: Time | 0.06874(0.1283) | 0.0480 |
| Age: Time | 0.016 (0.0442) | 0.0050 | | | |
| CTBPTB:Time | -0.217(0.0121) | 0.0023 | | | |
| HIVSneg:Time | -0.307 (0.2152) | 0.0130 | | | |
| dose: Time | -0.264 (0.1258) | 0.001 | | | |
| -2 log-likelihood =10240.04 | AIC=11175.428 | BIC=11376.1 | | | |

A joint model for the two outcomes of TB patients' weight and sputum status was fitted using adaptive Gaussian approximations that given the best mix of efficiency and accuracy. The model is resemble with the separate models except the sets of random intercepts and slopes for each response are now correlated rather than independent. It was fitted allowing for a linear time effect for each covariate that was selected as a fixed effect in the separate model. The subject specific random intercepts and random slopes were fitted to account for within-subject correlations. As shown table 11, all parameters are significant at 5% level of significance. Thus, the variable Sex, Age, Types of TB, HIV status, dose and time were identified as risk factors of body weight which are positively associated. But, age, types of TB, dose and time are significantly associated with positive sputum status. Except age, all are negatively associated with positive sputum status.

The parameter estimates and standard errors associated with the continuous outcomes are almost similar with the separate model shown in table 7. In some covariates, the standard errors are not exactly the same; the joint model is somewhat precise. Unlike the continuous parameters, the estimates for the binary outcome sputum status are a little bit different from the separate generalized linear mixed model which is summarized in table 10. Higher precision is observed for the estimates from the joint model (table 11).The resulting standard errors seem to suggest that there is higher precision in the joint model than the GLMM model. Therefore, the joint models tend to yield good precision than the separate analyses.

Based on the SAS PROC NLMIXED for joint model, the estimated variance covariance matrix and the estimated correlation matrix for random effects of both the weight and the sputum status as determined in the form of using equation (7) in section 3.5.3 have been shown in table 12 and table 13, respectively.

Table 12: The Variance-Covariance estimates for the joint model

| | | weight | | sputum status | |
|---|---|---|---|---|---|
| | | Intercept | slope | Intercept | slope |
| weight | Intercept | 10.283 | -2.135 | -2.722 | -0.961 |
| | slope | -2.135 | 1.998 | -0.924 | -0.628 |
| sputum | Intercept | -2.722 | -0.924 | 1.665 | 0.175 |
| | slope | -0.961 | -0.628 | -0.175 | 0.197 |

Table 13: The correlation matrix of the Random effect in the joint model

| | | weight | | sputum status | |
|---|---|---|---|---|---|
| | | Intercept | slope | Intercept | slope |
| weigh | Intercept | 1.00 | -0.471 | -0.658 | -0.659 |
| | slope | -0.471 | 1.000 | -0.649 | -0.698 |
| sputum | Intercept | -0.658 | -0.649 | 1.00 | -0.372 |
| | slope | -0.659 | -0.698 | -0.305 | 1.000 |

As shown table 12, the variability in the random intercept and slopes of weight is relatively higher than the variability of random intercept and slopes of sputum conversion. Similarly, the covariance between the random intercepts and slopes for weight is slightly extreme. The covariance's for both weight and sputum status are negative, which is indicative of a negative correlation, as seen in the correlation matrix (table13). The joint model used to investigate how the evolution of weight is associated with the evolution of sputum status, the association of the evolutions (AOE). The AOE has been determined by using equation (7) in section 3.5.3.Here the AOE between the random slopes for weight and the random slope for sputum conversion is -0.698. Thus, the negative value indicate a strong inversely association between the evolution of body weight and sputum conversion TB patients'.

### 4.4.1. Comparison of Joint and Separate Models

The separate models were fitted for the two outcomes together by assuming that ($\rho = 0$), which is entirely equivalent to fitting the models separately. Using joint model the evolution of these outcomes was investigated (table13).Hence, the body weight and sputum status show strong inverse relationship as evidenced by correlation of the random effects shown in the joint model ($\rho = -0.698$ (s.e. =0.134), p=0.0001) which is the correlation of the two random slops and also the correlation between the random intercepts clearly show a negative strong association between them (table13). The Comparison of parameter estimates of the separate and joint models have been shown below table 14.

Table 14: Parameter estimates and standard errors for separate and joint model

| | Joint model | | Separate Model | |
|---|---|---|---|---|
| **Effects** | Estimates(s.e) | p-value | Estimate(S.e) | p-value |
| (Intercept) | 29.709(1.4514) | 0.0000 | 29.689(1.462) | 0.0000 |
| Sexmale | 6.176(0.7041) | 0.0004 | 6.166(0.714) | 0.0000 |
| Age | 0.188 (0.003) | 0.0166 | 0.187(0.006) | 0.0000 |
| CTBPTB | 0.142 (0.0158) | 0.0004 | 0.130(0.0743) | 0.0073 |
| HIVSneg | 0.798 (0.0655) | 0.0001 | 0.796(0.0774) | 0.0057 |
| dose | 3.816(0.2703) | 0.0141 | 3.727(0.432) | 0.0000 |
| Time | 1.886(0.04133) | 0.0001 | 1.782(0.498) | 0.0022 |
| Sexmale: Time | -0.179(0.0776) | 0.0049 | -0.1807(0.198) | 0.0085 |
| Age:Time | 0.016 (0.0442) | 0.0050 | 0.015(0.0487) | 0.0090 |
| TBCPTB:Time | -0.217(0.0121) | 0.0023 | -0.216(0.186) | 0.0015 |
| HIVSneg:Time | -0.307 (0.2152) | 0.0130 | -0.305(0.257) | 0.0045 |
| Dose: Time | -0.264(0.137) | 0.001 | -0.252(0.137) | 0.002 |
| RE.Var($b_{10}$) | 10.283(1.1403) | 0.003 | 8.284 (1.888) | 0.0087 |
| RE.Var($b_{11}$) | 1.998(0.4272) | 0.0045 | 0.565(0.813) | 0.0087 |
| $\sigma_1^2$ | 2.718 (2.112) | 0.0007 | 2.177( 2.228) | 0.0015 |
| Binary Outcome (Sputum ) | | | | |
| Effects | Estimates (S.e) | p-value | Estimate(S.e) | p-value |
| intercept | 1.2718 (0.07269) | 0.0043 | 1.2097(0.06460) | 0.00494 |
| Age | 0.03027 (0.0297) | 0.0075 | 0.0137 (0.2526) | 0.0001 |
| CTBPTB | -2.8508 (0.0255) | 0.0001 | -2.604(0.1865) | 0.0058 |
| dose | -0.1907(0.04309) | 0.0154 | -0.1704 (0.7464) | 0.0067 |
| Time | -2.553 (0.0420) | 0.0018 | -2.022(0.0131) | 0.0080 |
| Time: Age | 0.096(0.2107) | 0.0058 | 0.0018 (0.3112) | 0.0001 |
| Time: CTB | 2.4947(0.0217) | 0.0001 | 1.3271(0.2112) | 0.002 |
| dose: Time | 0.06874(0.1283) | 0.0480 | 0.0188 (0.1362) | 0.0032 |
| RE.Var ($b_{20}$) | 1.665(0.1272) | 0.0154 | 0.849 (1.010) | 0.0712 |
| RE.Var ($b_{21}$) | 0.197 (0.034) | 0.0001 | 0.080 (0.914) | 0.0032 |
| $\sigma_2^2$ | 2.969 (0.0446) | 0.0494 | 1.2697(0.065) | 0.0029 |
| Common parameters | | | | |
| Corr.RE ρ | -0.698(0.134) | 0.0001 | - | - |
| −2LLL | AIC | | -2LLL | AIC |
| 10240.04 | 11175.43 | | 10304.17 | 11424.43 |

RE=random effect s.e= standard error, LLL=loglikeliood, AIC=Akaike's information criterion

When comparing the results from the independent setting to the results from the multivariate setting (Table14), the likelihood comparison shows a convincing improvement in model fit, when random effects are allowed to correlate. Comparing the separate and joint models, although parameter estimates for both outcomes are nearly equivalent, small changes are observed in parameter of some covariate. In order to decide the best model, from the fitted model the corresponding loglikeliood values were obtained. The Loglikeliood value of the joint model is $LR_{joint} = -2loglikelihood = 10204.06$ and the separate models were fitted for two outcomes together by ignoring the correlation between them (i.e.$\rho = 0$) which entirely equivalent to fitting the two independent models separately as results were shown in Table14, but a single likelihood value has obtained that enables to comparison with the correlated random effect model (joint) (i.e. $LR_{sep} = -2\log likelihood = 10304.17$ ). The asymptotic null distribution of the likelihood ratio test statistic is a chi-square distribution with degrees of freedom equal to the difference between the numbers of parameters in the two models. Hence, the joint model of the two outcomes weight and sputum conversion is significantly better than two separate models ($\chi^2$ =116.13, df=4, P-value< 0.0001). Another point to touch upon is how the covariates compare between the two types of models. From table 10, both the separate and joint models did find significant relationships between baseline types of TB with weight. The estimate was increased from 0.142 to 0.130; as result, the standard error has declined from (0.0743 to 0.0158, p=0.0004) which is smaller for the joint model. Similarly, it has associated with the sputum status with estimate increased from -2.604 to -2.851 and standard error decreased from (0.0865 to 0.0255, p=0.0001).Both models drew the same conclusions with regards to the rest covariates being related to with weight and sputum conversion, with the joint model, in general, having more precise estimates, as shown table 14, by smaller standard errors.

Table 15: Covariance estimates and correlation estimates for separate and joint model

| | **Covariance estimates** | | **Correlation estimates** | |
|---|---|---|---|---|
| Effects | Separate Estimates (s.e) | Joint Estimates (s.e) | Separate Estimates (s.e) | Joint Estimates (s.e) |
| $\sigma^2_{b10}$ | 8.284(1.888) | 10.283(1.140) | 1.000 | 1.000 |
| $\sigma^2_{b11}$ | 0.5265(0.813) | 1.998(0.427) | 1.000 | 1.000 |
| $\sigma_{b10,b11}$ | -0.565(0.401) | -2.135(0.121) | -0.212 | -0.471 |
| $\sigma^2_{b20}$ | 0.849 (1.010) | 1.665(0.127) | 1.000 | 1.000 |
| $\sigma^2_{b21}$ | 0.080 (0.914) | 0.197(0.034) | 1.000 | 1.000 |
| $\sigma_{b10,b20}$ | - | -2.722 (0.061) | - | -0.658 |
| $\sigma_{b20,b21}$ | -0.202(0.724) | -0. 175(0.404) | -0.305 | -0.372 |
| $\sigma_{b10,b21}$ | - | -0.961(0.133) | - | -0.659 |
| $\sigma_{b11,b20}$ | - | -0.924 (0.227) | - | -0.649 |
| $\sigma_{b11,b21}$ | - | -0.628(0.017) | - | -0.698 |

Like parameter estimation and testing of the fixed effect, the random effect is another important aspect. High variability is the indicator of less accuracy or high error on prediction of the association of outcome evolutions with respective risk factors. As shown in table15, the standard errors of the subject specific random intercept and random slope of the joint model are slightly smaller, when compared to the separate models. The SE's are further evidence that fitting a joint model is a better method.

## 4.5. Discussion

In this paper, joint model for the association of longitudinal binary and continuous processes has been developed to see the joint evolution of sputum status and weight variation. Papers taking this type of approach include Catalano and Ryan (1992) and Regan.et al. (1999). The former approach is the latent variables approach where as the later is the correlated random effect approach. Thereby, the joint model directly related the later approach that shows association of these longitudinal outcomes by incorporating correlated subject specific effects.

Consequently, data exploration for continuous data, separate parameter estimates a linear mixed model (LMM) and generalized linear mixed model (GLMM) and joint model were fitted. From individuals profile plot, both within and between individuals variability of the weight of TB patients were existed .The mean evolution also indicated that the body weight has increased linear pattern over time. From this evolution, the mean weight of male HIV negative pulmonary tuberculosis patients is higher than female HIV negative pulmonary tuberculosis patient.

From a separate LMM, first the importance of both random intercept and slope was done using AIC, BIC and likelihood ration test, in doing so the linear effect of time also checked through these criterions. Then, the data were analyzed using different linear mixed model incorporating patient specific weight variability. The saturated parameter estimation model was selected that supports a significant assumption of homogeneous variances of subject specific variability of patients. Also, different variance covariance structures were done to select the best variance covariance structure for the best model, thus linear mixed model with linear time effect and unstructured variance components with uncorrelated and identically distributed error term was fitted well to the weight of TB patients. All covariates and their interaction with time in the selected model have a significant effect on the body weight variation of tuberculosis patients. This finding regarding the positive association of weight and the covariates sex, category of TB, dose and age coincides with previous studies in Ethiopia (Hiwot A.et al. 2013). Age with time interaction has significant effect on the body weight supported by Rios J.et al. (2011) those found that after the patients started TB diagnosis their body weight increases. Similarly, generalized linear mixed model (GLMM)

was employed for analysis of the binary longitudinal sputum status and except BIC, similar model section criterions were used to select best linear time effect on the sputum. All the base line predictor variables under the selected model were significantly important on the sputum conversion. The age of patients has a negative effect which is supported by Worodria et al. (2011) and C-S Wang. (2008) the older age tuberculosis patients increased risk of having a positive sputum conversion. The level of dose that was provided during diagnosis minimized the positive sputum conversion of patients which the same with the previous study done by Xuefeng Liu and Michael J. Daniels (2003). After the separate analyses of each data, the joint model which necessitates the modeling of associations between the outcomes of at the same time point and to take the precise parameter estimation model. These two sub-models were linked via correlated random effect. Specifically, the longitudinal sub models were described by both the usual linear mixed model and generalized linear mixed model incorporating subject specific variances, which is consistent with previous findings of Gueorguieva, R.V., and Sanacora (2006).This is accomplished with the incorporation of correlated random effects (i.e., random intercepts and random slopes) in the individual linear mixed model and generalized linear mixed model for the outcomes. In fact, a joint longitudinal model for a binary and continuous outcome measured over time. The aim of the joint model was to study the relation between body weight and sputum status. The associations between the evolutions between the two outcomes were investigated. Results of the joint model suggested a very strong association that confirms similar findings from Hoa, N.B.et al. (2012). It was important to use joint model methods for inference in order to avoid biased results supported by Regan.et.al. (1999) this was achieved by incorporating the model of Heagerty (1999) into the correlated random effects joint models used to jointly model two longitudinal outcomes. In the selection of methods for inference, the maximum likelihood method generally produced most reliable results and adaptive Gaussian Quadrature (nAQG=20) was used. The selection of saturated joint model was facilitated by AIC value that has been compared among various realistic models. It has been used for determining the random effects to be included in the longitudinal model and to select the best joint model among several candidate models. In terms of implementation, the employed method allows to efficiently make use of available resources, such as the SAS procedure NLMIXED. However, the saturated model did not obtain easily. A limitation of the joint model was its

intensive computation. The analysis required approximately 3-4 days with 100 iterations. This makes it difficult to use the joint model as a real-time.

# CHAPTER -FIVE

# CONCLUSION AND RECOMMENDATION

## 5.1. Conclusion

This study closely examined separate and joint models for binary and continuous outcomes via correlated random effect for data obtained from Jimma University specialized Hospital TB clinic. Separate analysis of the weight change and sputum status proved that incorporation of linear time random effects in the fitted model. Specifically, the assumption of homogeneous also checked for both outcomes. All the covariates; Sex, Age, category of TB, dose and HIV status that were included in the linear mixed model are significantly associated with weight change and similarly all covariates that were included in the best model; Age, dose and category of TB are significantly associated with the sputum status of tuberculosis patients.

Turning to the analyzed joint model, where body weight as well as sputum status were measured repeatedly for each patient. The two outcomes were modeled jointly to capture association between them. The two end points show a strong inverse relationship as evidenced by the correlation of the random effects; such that a patient with negative sputum result has higher weight than positive sputum result of TB patients'. Furthermore, model fit was improved when random effects are allowed to correlate. Comparing the separate and joint models, while parameter estimates for the continuous outcome remain the same, small changes are observed in the binary part. As a result, a joint modeling approach tends to provide unbiased and more precise estimates.

## 5.2. Recommendation

According to the findings different risk factors were identified that influence the body weight and sputum conversion of TB patients. In this thesis, sex, age, HIV status, dose, types of TB and time are associated with body weight and except sex and HIV status the rest are risk factors for sputum conversion. As a result, negative HIV status patients had higher body weight than that of positive HIV status. Thus, special attention should be given to HIV

positive patients during TB diagnosis to extend their lives. Similarly, the level of dose has associated with sputum status and body weight, health workers should be cautious when medicine or dose has been provided. Even if, weight loss and sputum status are symptom of tuberculosis, it is also a curable disease if properly treated, attention must be given to DOTS strategy by government and non-government organizations. However, these are not enough to evaluate body weight and sputum status of patients' over time depending on TB diagnosis; also it is important to know factors that influence patients' bodyweight and sputum conversion. Hence, Further studies are also recommended with additional exogenous variables (such as smoking status, income, duration of diabetes, marital status, alcoholism and educational status...etc.) to see the progression of body weight and sputum conversion of tuberculosis and to strengthen and explore the problem among TB patients in depth with large sample sizes and advanced diagnostic techniques.

A joint model has been performed to assess the evolutions of the outcomes over time by considering correlation between them. The resulting model showed association between the two responses and yields parameter estimates which were close to those from single-outcome analyses but provided higher precision. The difference in precision could affect inferences. Thus, it is important to make use of such joint modeling approaches, which tend to provide unbiased and more precise estimates.

**Limitations of the Study**

This thesis used a retrospective study. The data collection was based on the available clinical records and on the data available in the tuberculosis registers; hence, the patients' history didn't register well. Important covariates like smoking status, diabetes mellitus, marital status, alcoholism and educational status were not included in the TB register. In fact, this study could not investigate the influence of these variables on the body weight and sputum conversion of TB patients. In addition, shortage of related research also one constraint that influences to realize this thesis.

# References


Abramowitz and Stegun I. A. (1964).Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. Dover, New York.

Antonio K. and Beirlant J. (2006).Actuarial Statistics with Generalized Linear Mixed Models, University Center for Statistics, Belgium.

Becerra MC.et al. (2000).Using treatment failure under effective directly observed short-course chemotherapy programs to identify patients with multidrug-resistant tuberculosis. Int J Tuberc Lung Dis **4**: 108–114.

Brudney K. and Dobkin J. (1991).Resurgent tuberculosis in New York City. Human immunodeficiency virus, homelessness, and the decline of tuberculosis control programs. Am Rev Respir Dis; **144:745**-9.Care Med **174**: 344–348.

Catalano, P.J., and Ryan, L.M.(1992).Bivariate latent variable models for clustered discrete and continuous outcomes. Journal of the American Statistical Association, **87**, 651-658.

Compoux JJ, et al. (2004). Sheris medical microbiology. In: Ryan KJ, Ray CG, editors. Pathogenic bacteria: mycobacteria. London: McGraw-Hill ,443-451.

Daniels, M.J., and Normand S-L (2006). Longitudinal profiling of health care units based on continuous and discrete patient outcomes. Biostatistics, **7**, 1-15.

Diggle et al. (2002). Analysis of Longitudinal Data. New York, NY: Oxford University Press.

Davis, et al. (2002). Statistical Methods for the Analysis of Repeated Measurements. New York, NY: Springer.

Dunson, D.B. (2003). Dynamic latent trait models for multidimensional longitudinal data. Journal of the American Statistical Association, **98**, 555-563.

England A.et al. (2003).Body mass index in adolescence in relation to total mortality: 32-year follow-up of 227,000 Norwegian boys and girls. Am J Epidemiol; **157(6):** 517- 23.

Fitzmaurice GM, Laird NM andWare JH (2004). Applied Longitudinal Analysis; Wiley-Interscience, editor. New Jersey: John Wiley & Sons, Inc.

Fitzmaurice, G. M. and Laird, N. M. (1993) .A likelihood based method for analyzing longitudinal binary responses. Biometrika. **80**, 141–151.

Gueorguieva, R. (2013). Random effects models for joint analysis of repeatedly measured discrete and continuous outcomes. In: Analysis of Mixed Data: Methods & Applications .

Gueorguieva, R.V., and Agresti, A. (2001). A correlated probit model for joint modeling of clustered binary and continuous responses. Journal of the American Statistical Association, **96**, 1102-1112.

Hardin J and Hilbe JM (2003). Generalized estimating equations; Hall/CRC C, editor. Washington DC: CRC Press Company.

Heagerty, P. J. (1999). Marginally specified logistic-normal models for longitudinal binary data. Biometrics, **55**:688–698.

Hedeker, D., & Gibbons, R.D. (2006). Longitudinal Data Analysis. Hoboken, NJ: John Wiley & Sons, Inc.

Henderson, C. (1984). Best linear unbiased prediction (BLUP) and restricted maximum likelihood (REML) estimation of parameters of Gaussian linear mixed models.

Hirakawa, A. (2012). An adaptive dose-finding approach for correlated bivariate binary and continuous outcomes in phase I oncology trials.

Hiwot A.et al.(2013).Smear positive pulmonary tuberculosis among diabetic patients at the Dessie referral hospital, Northeast Ethiopia. Infectious Diseases of Poverty ,2:6 doi:10.1186/2049-9957-2-6

Hoa, N.B., et al.(2012). Changes in Body Weight and Tuberculosis Treatment Outcome in Vietnam. The International Journal of Tuberculosis and Lung Disease, **17**, 61-66. http://dx.doi.org/10.5588/ijtld.12.0369

Issar S (2003). Mycobacterium tuberculosis Pathogenesis and Molecular Determinants of Virulence.

Jones Lopez et al. (2011). Effectiveness of the standard WHO recommended retreatment regimen (category II) for tuberculosis in Kampala, Uganda: a prospective cohort study. PLoS Medicine 8: e1000427.

Laird N. and Ware J. (1982). Random-effects models for longitudinal data. Biometrics.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. Biometrika, **73**:13–22.

McCulloch, C.E. (1994). Maximum likelihood variance components estimation for binary data. Journal of the American Statistical Association **89**, 330-335 Medicine.330:17031709.

Ministry of Health of Ethiopia ( 2008): Tuberculosis, Leprosy and TB/HIV Prevention and Control Programme Manual. 4th edition. Addis Ababa:

Molenberghs G. and Verbeke, G. (2005). Models for Discrete Longitudinal Data. New York: Springer.

Olkin I. and R. F. Tate (1961). Multivariate correlation models with mixed discrete and continuous variables. Annals of Mathematical Statistics 32, 448{465. (with correction in **36**, 343(344).

Ottmani S, et al. (2006).Results of cohort analysis by category of tuberculosis re-treatment cases in Morocco from 1996 to 2003. International Journal of Tuberculosis And Lung Disease. **10**:1367-1372.

Pinheiro et al. (2010).Linear and Nonlinear Mixed Effects Models. R package version **3**.1-97.

Pinheiro, J. C. and Bates, D. M. (2000). Mixed-Effects Models in S an S-PLUS, Springer

Regan, et al.(1999). Likelihood models for clustered binary and continuous outcomes: Application to developmental toxicology.

Rios J.et al. (2011).Weight Variation over Time and Its Association with Tuberculosis Treatment Outcome: A Longitudinal Analysis, Universidad Peruana Cayetano Heredia, Peru,

Skrondal, A. and Rabe-Hesketh, S. (2008). Multilevel and related models for longitudinal data. In J. d. L. E. Meijer (Ed.), Handbook of Multilevel Analysis (pp. 275-299). New York, NY: Spring Science + Business Media.

Smith I.(2003).Mycobacterium tuberculosis pathogenesis and molecular determinants of virulence.ClinMicrobiolRev, **16(3)**:463-496

Tabrasi et al.(2008). Early initiation of anti-retroviral therapy results indecreased morbidity and mortality among patients with TB and HIV. Journal of International AIDS Society 12:14, dio 10.1186/1758-2652- 12.14

Tsiatis et al. (1995). Modeling the relationship of survival and longitudinal data measured with error, Application to survival and CD4 counts in patients with AIDS. Journal of the American Statistical Association, **90**, 27–37.

Van Crevel R.et al.(2002). Decreased plasma leptin concentrations in Tuberculosis patients are associated with wasting and inflammation. J Clin Endocrinol Metab; **87(2)**:758-63.

Vonesh, E. F., & Chinchilli, V. M. (1997). Linear and nonlinear models for the analysis of repeated measurements. New York, NY: Marcel Dekker, Inc.

WHO (2005): Global tuberculosis control. Geneva: WHO report; WHO/HTM/TB/2005.349

WHO (2008): Global tuberculosis control: Surveillance, planning and financing. Geneva: WHO/HTM/ TB/2008.393

WHO (2009): Global tuberculosis control-epidemiology, strategy, financing. WHO/HTM/TB/2009.411

WHO (2010): Global tuberculosis control. Report No. 2010.7

WHO (2011): Global tuberculosis control.WHO/HTM/TB/2011.16

World Health Organization (2006)., Anti-tuberculosis treatment in children. Geneva, Switzerland. International Journal of Tuberculosis and Lung Diseases, **10(11)**:1205-1211.

Worodria et al (2011).“Nucleic acid amplification tests for diagnosis of smear-positive T B: a prospective cohort study,” PloS One , vol. **6**, no.1, article e16321, 2011.

Wulfsohn and Tsiatis (1997). A joint model for survival and longitudinal data measured with error. Biometrics, **53**, 330–339.

Xuefeng Liu and Michael J. Daniels (2003).Joint Models for the Association of Longitudinal Binary and Continuous Processes with Application to a Smoking Cessation Trial, Brown University.

Yohannes et al. (2013). Smear positive pulmonary tuberculosis disease at University of Gondar Hospital, Northwest Ethiopia. BMC Research, 6:21 doi:10.1186/1756-0500-6-21

# Appendix -1: Model diagnosis for Linear Mixed Model

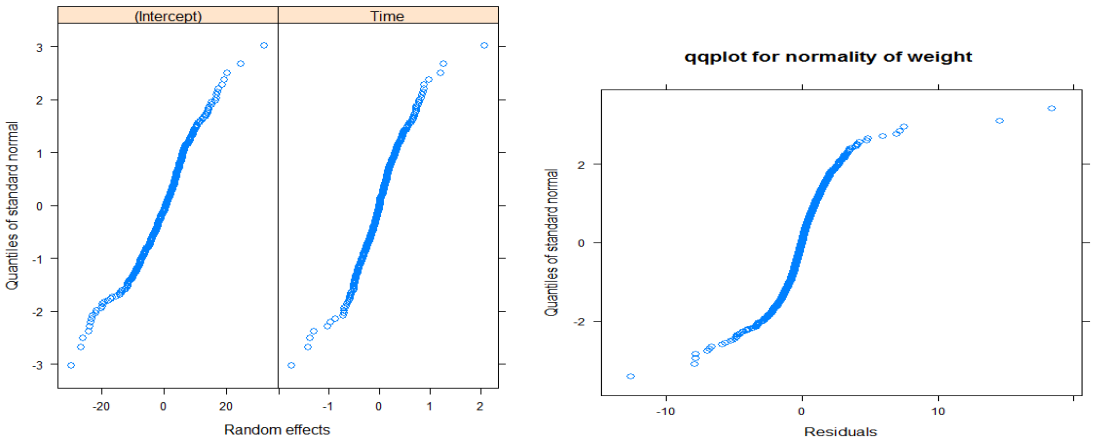



Figure 4.23: Q-Q plots for random intercept and slopes

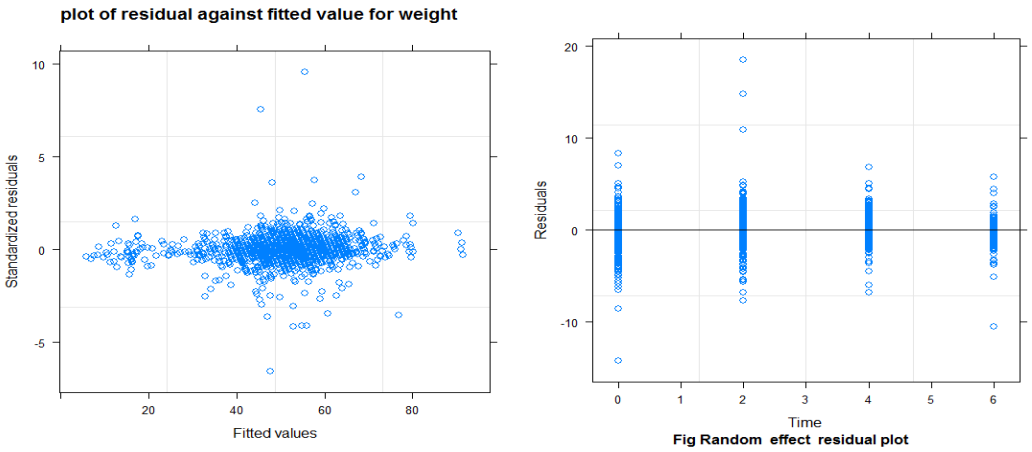



**Figure 4.24** Residuals vs fitted value

# Appendex-2: Model diagnosis for Generalized Linear Mixed Model

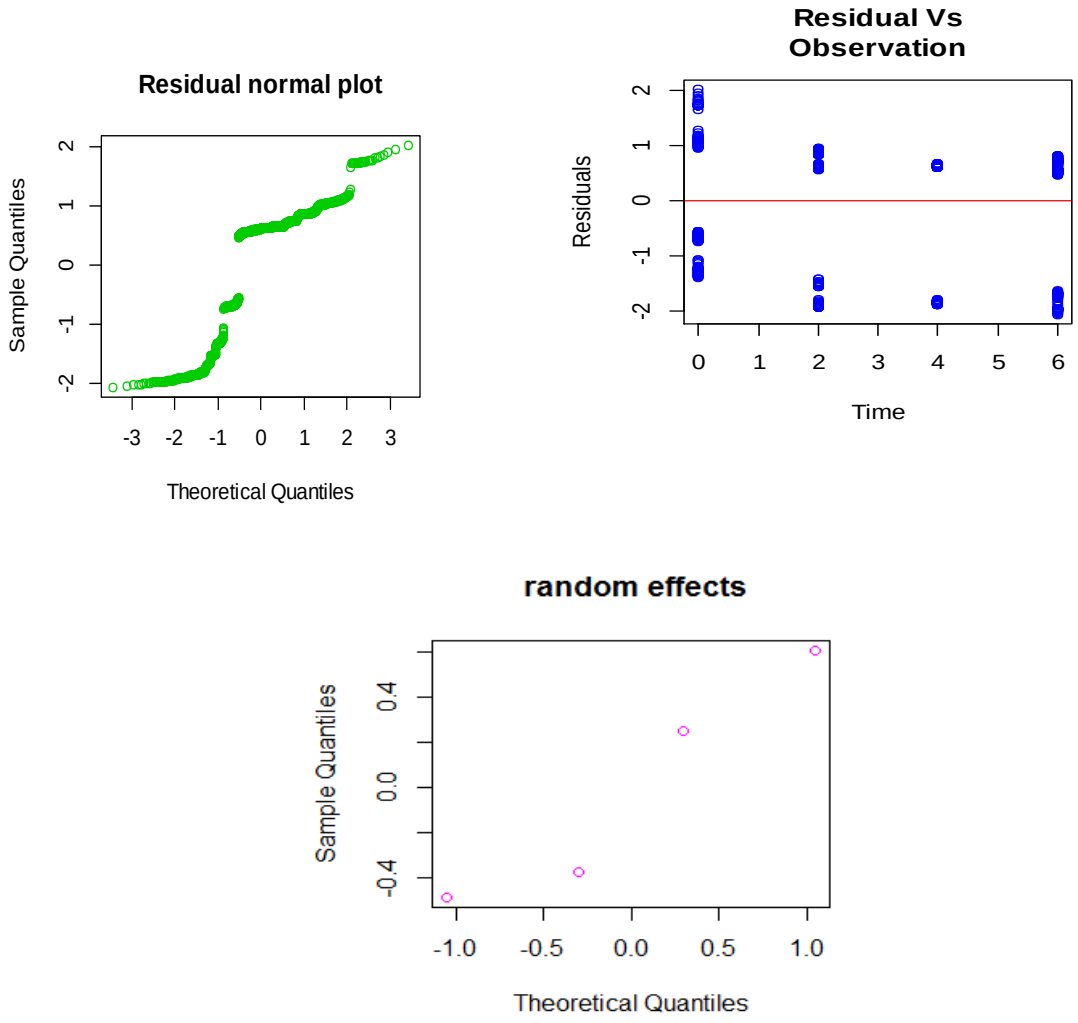



Figure 4.25 Diagnosis plots for the generalized linear mixed model