Joint Generalized Nonlinear Mixed Models for Longitudinal Data

by

Ashiqul Haque

MS (2018), BSc (2017), Shahjalal University of Science and Technology, Bangladesh

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

In the Graduate Academic Unit of Mathematics and Statistics

Supervisor(s):	Tariq Hasan, Ph.D, Statistics
	Renjun Ma, Ph.D, Statistics
Examining Board:	Guohua Yan, Ph.D, Statistics, Chair
	Jeffrey D Picka, Ph.D, Statistics
	Murshed Chowdhury, Ph.D, Economics

This thesis is accepted by the
Dean of Graduate Studies

THE UNIVERSITY OF NEW BRUNSWICK

August, 2020

Abstract

Joint modeling of multiple longitudinal responses enables us to account for the association between them and is thus more efficient than seperate analyses. Most existing techniques to handle this problem are based on the assumptions of normality of the responses and linearity of the mean functions. However, non-normality of responses and non-linear shape of their mean functions often arise from medical and population growth studies. For example, it is desirable to investigate the nonlinear mean structures in the analysis of the effect of different drug formulations while accounting for their association in Pharmacodynamics (the study of what the drug does to the body). We propose to model data of mixed types jointly by incorporating both subject-specific and time-specific random effects into Tweedie nonlinear models. An optimal estimation procedure for our model has been developed using the orthodox best linear unbiased predictors of the random effects. This approach allows us to model multiple non-normal longitudinal responses with interpretable parameters.

Dedication

This thesis is dedicated to all the healthcare professionals around the world who are working hard at the frontline to provide the much needed services for the COVID-19 patients.

Acknowledgements

I have an enormous amount of gratitude towards my supervisor Dr.Tariq Hasan for his constant guideline and support both academically and financially. I can't imagine the completion of this thesis without his patient mentorship. I am grateful to my another supervisor Dr.Renjun Ma for all his crucial suggestions that shaped the direction of my work to a smooth end. I would like to express my sincere gratitude to Dr.Guohua Yan and Dr.Jeffery Picka for allowing me to take their extraordinary classes and supervising my seminar talks. I am also grateful to Dr.Murshed Chowdhury and Dr.Rafiqul Islam Chowdhury for their remarkable academic mentorship that helped me to go through my academic journey at the University of New Brunswick.

I am thankful to all the faculty members and staffs at the Department of Mathematics and Statistics for an excellent learning environment that I have never experienced before.

Finally I would like to thank my family, lab mates, and friends for their love and support. It means a lot to me.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In longitudinal studies, measurements are taken on the same subjects of interest at different timepoints. Longitudinal studies mainly focus on the mechanism/process that takes place within an individual and its variation across the population. In order to make inferences about such mechanisms, repeated measurements of the outcome(s) on each of the sample subjects from the population are recorded. Both cross-sectional (one time point, many subjects) and time-series (one subject, many time points) data can be seen as special cases of longitudinal data. Datasets arising from longitudinal studies have some unique features that make it difficult to adopt the methods available for other types of data analysis. For example, the independence assumption which is common in statistical models is not satisfied here, since the measurements collected from the same subject on several occasions cannot be independent in general. The most attractive feature of any longitudinal study is its capacity to explore the change in a variable over time in a meaningful way.

Longitudinal data analysis plays a key role in developing models to identify the pattern of changes in response variable and/or establishing relationships between outcome and covariates. Nonlinear change in response over time arises

in many areas including Biomedical research, Population growth, and Econometrics (Deisboeck and Kresh, 2007; Sabzpoushan, 2020; Lucia et al., 2014). Particularly this type of change commonly occurs in the study of Pharmacokinetics (what the body does to the drug) and Pharmacodynamics (what the drug does to the body). To illustrate, concentration responses in blood from different formulations of drugs designed for the treatment of hypertension change nonlinearly over time (Willemsen et al., 2017). Moreover, virologic suppression and CD4 cell counts in blood from people living with HIV show a nonlinear recovery pattern after initiation of antiviral treatments (Dronda et al., 2002). Building non-linear mixed-effects models (NLMM) is the most widely used approach to model the longitudinal responses having nonlinear changing patterns (Pillai et al., 2005). This method enables researchers to detect the systematic and random variation within and between individuals under study. This framework has been going through significant development of new methodological and computational techniques since the 1990s. Davidian and Giltinan developed a rich overview of the construction, interpretation, and application of nonlinear mixed-effects models (Davidian and Giltinan, 2003). Relying on normality and homogeneity assumption about the longitudinal response is a common practice in NLMM formulation. Following such a procedure may result in unreliable parameter estimates in situations when skewness and heterogeneity arises. Several studies analyzed datasets constructed from HIV research by transforming responses into normal. Similar approaches were taken to deal with censoring in the response and measurement error in covariates (Wu and Ding, 1999; Wu, 2004). Lu and Huang proposed a Bayesian approach to nonlinear mixed-effects models for longitudinal data with heterogeneity and skewness (Lu and Huang, 2014). In another study, the authors proposed a Bayesian approach to find unknown parameter estimates and random effects from nonlinear

reproductive dispersion mixed-effects model (Tang and Zhao, 2014).

In order to analyze data from longitudinal studies that have multiple responses measured on different occasions, researchers may need sophisticated techniques to develop the required models. One potential way to analyze such datasets is to model the outcomes separately. This approach is helpful to assess the influence of the treatment (intervention) on each response. Advanced techniques, particularly joint models come into play when researchers attempt to answer complex questions like, how are the outcome-specific evolutions related to each other(Fitzmaurice and Ravichandran, 2008)? This type of model can also help to get an overall idea about the heterogeneity among the subjects under study. Joint models for longitudinal and survival data are the most widely used joint model. Like most other statistical modeling approaches, the use of linear joint models for longitudinal data is relatively common compared to their non-linear versions. But linear models often come with the loss of straightforward interpretability of the estimated parameters.

Proust-Lim et al have developed a nonlinear joint modeling approach for multivariate longitudinal data (Proust-Lima et al., 2007). Hu and Sale (Hu and Sale, 2003) have developed an extension of linear and generalized linear models for responses with informative dropouts to nonlinear models. Joint nonlinear mixed-effects models can be fitted to analyze longitudinal image data from early brain development studies generating Magnetic resonance imaging (MRI) reports (Wolff et al., 2017). The development of joint models following the Bayesian approach is also evident in the recent statistical literature. Semiparametric joint NLMM and a finite mixture of NLMM have been shown to perform well when data contain skewness, measurement errors, and missing covariates (Huang and Dagne, 2012; Lu et al., 2016).

Joint models for more than one response can be formulated in multiple ways.

For instance, when all the outcomes under consideration are of the same types (such as continuous), joint models can be constructed by specifying a joint density directly. Conditional models can be developed by defining models for each of the outcomes separately. Another frequently used joint modeling approach for longitudinal data is shared parameter models or random-effects models. Usually, this method builds a model based on the assumptions about unobserved variables (random-effects) shared by the outcomes to be modeled together. The core assumption behind such models is the conditional independence between outcomes given the random-effects (Fitzmaurice and Ravichandran, 2008).

In most cases, existing models can accommodate responses from a limited number of distributions such as Normal (for continuous) and Poisson (for the count). Besides, most of the available shared parameter models require the random effects to be normally distributed and usually applicable to linear models only (Fitzmaurice and Ravichandran, 2008). It is desirable to develop flexible shared parameter models that can accommodate non-normal random effects and applicable to both linear and nonlinear models.

This thesis proposes joint models for the longitudinal responses belonging to the Tweedie exponential family with flexible distributional assumptions about random effects. The foundation of this modeling framework lies in models for responses from Tweedie exponential distributions (Jørgensen, 1987; Ma, 1999). The proposed models incorporate two levels of random effects (subject-specific and time-specific) to capture the within and between-subject correlation structures present in the repeated measures. An optimal estimation process for our model has been established using the orthodox best linear unbiased predictors of the random effects. The proposed nonlinear joint models are applicable to both balanced and unbalanced longitudinal datasets.

The rest of the chapters of this thesis are organized as follows,

The next chapter reviews the available statistical analysis techniques for longitudinal data. Different modeling approaches, mean functions, and covariance structures are described. The Tweedie exponential family of distributions is introduced in short. Several joint model construction strategies are also discussed.

The proposed model based on three key assumptions about the random-effects and responses is introduced in Chapter 3. The marginal moments of the responses and the random effects are derived from the conditional and unconditional moments of the random effects and the responses as per necessity. The estimation process of the regression parameters, the predictors of the random effects, and the correlation parameters are described in this chapter.

The application of the proposed model is shown by analyzing a Losartan (an anti-hypertensive drug) dataset in Chapter 4. The concentration level of two comparable drug formulations in blood are modeled together by following the joint model framework.

The results from separate univariate models for the response variables under consideration are also included to compare and contrast the results from the proposed joint models.

Chapter 5 presents an evaluation of our proposed data analysis technique from the the results of well designed simulation experiments.

Finally, a conclusion about the proposed model, including discussion and recommendation for potential future works in this line of research is provided in Chapter 6.

Chapter 2

Review of Literature

In this chapter, the available statistical models used to analyze longitudinal data are reviewed, along with various data structures that can arise in longitudinal studies.

2.1 Longitudinal study

The fundamental idea behind longitudinal study design is the collection of measurements on the same individuals/subjects of interest repeatedly over time. Longitudinal studies are common in the fields of medicine, social-personality, and clinical psychology. This type of research can play an important role in establishing a causal relationship between variables (Aschengrau and Seage, 2013). For example, if a researcher wants to identify which disease affects a specific group of people in the population, then observing that group of individuals over some time to collect meaningful data will be helpful to answer the researcher's question of interest. Longitudinal studies are not just restricted to the field of science or medicine; they have a considerable influence in the area of business as well (Linnhoff et al., 2020). With a longitudinal study, one can measure and compare various business and branding aspects by deploying surveys

focusing on market trends, product feedback, and customer satisfaction.

A basic structure for a longitudinal dataset with a single response can be displayed in the following form,

Table 2.1: **Longitudinal data structure**

Subject	Measurement occasion	Response, Y_{it_i}	Covariates, X_{it_im}
1	1	Y_{11}	$x_{111} \dots x_{11m}$
$\dots$	$\vdots$	$\vdots$	$\vdots$
1	t_1	Y_{1t_1}	$x_{1t_11} \dots x_{1t_1m}$
2	1	Y_{21}	$x_{211} \dots x_{21m}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
2	t_2	Y_{2t_2}	$x_{2t_21} \dots x_{2t_2m}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
I	1	Y_{It_1}	$x_{I11} \dots x_{It_Im}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
I	t_I	Y_{It_I}	$x_{It_I1} \dots x_{It_Im}$

where Y_{it_i} represents the measurement on the response Y for the i-th subject at the t-th occasion,

X_{it_im} denotes the m-th covariate value for the i-th subject at the t-th occasion.

Longitudinal studies can be classified into different types based on specific designs and data structures (Fitzmaurice and Ravichandran, 2008). Usually, longitudinal study designs aim at gathering a fixed number of repeated measurements on all the subjects under study. The study is said to be "balanced" when an equal number of repeated measures collected at a common set of time points are available for all the participating individuals. On the other hand, in some situations, studies may end up with getting an unequal number of observations and/or mistimed (collected at a different set of time points) observations for the participants. In such cases, the resulting dataset is described as "unbalanced". From another point of view, when researchers look

back in time using an existing database (such as medical records or insurance claims) longitudinal studies can be "retrospective" (Mathalon et al., 2000). In contrast, studies requiring the collection of new data are called "prospective" (Rosen et al., 2017).

In most cases, longitudinal studies are observational, but in certain situations, they can also be designed as randomized experiments. One of the major advantages of longitudinal studies over cross-sectional studies is the assessment of within-subject changes in the response over time. Different longitudinal data analysis techniques can be utilized to make conclusion about the pattern of changes in the response variable both at the population level and the individual level of observational units (subjects) (Fitzmaurice and Ravichandran, 2008).

Typically longitudinal studies are time-consuming and expensive (Aschengrau and Seage, 2013). Besides, longitudinal studies can suffer heavily from missing observations arising from loss to follow-up.

2.2 Longitudinal data analysis

In its simplest form, longitudinal data analysis may consist of paired sample t-test (Weiss, 2005b) and Multivariate Analysis of Variance (Taris, 2000) in the context of non-experimental longitudinal survey research. Over the last three decades, data analysis techniques for longitudinal studies have become remarkably sophisticated (Grimm, 2019). Singer et al and Fitzmaurice et al introduced several approaches for analyzing repeated measures are (Singer et al., 2003; Fitzmaurice et al., 2012). A conceptually straightforward way to analyze data from a balanced longitudinal study with discrete covariates (such as Treatments or Interventions) is the analysis of response profiles (Fitzmaurice

et al., 2012). One of the key features of the analysis of response profiles is that it allows for arbitrary mean and covariance patterns. Even though the application of this technique requires the data to come from a balanced design, it can handle incomplete data due to limited missing observations in the response. A more advanced approach to deal with longitudinal data is to fit a parametric or semi-parametric curve. In situations where the longitudinal data are essentially unbalanced over time, fitting a parametric curve with models for the mean response is an appealing approach to follow (Fitzmaurice et al., 2012).

2.2.1 Modeling approaches

Two broad classes of models for longitudinal data are marginal models and conditional models (Diggle et al., 2002).

In marginal models, regression of the response and within-subject correlation are modeled in two different parts. The regression part considers the marginal expectation to be a function of explanatory variables. The within-subject correlation is assumed to be a function of the marginal means and possibly of additional parameters.

Let Y_{ij} and μ_{ij} represent the measurements on a response Y and the average of the response respectively for i-th subject at j-th level of a covariate X at different timepoints. If ρ denotes the within-subject correlation between two levels of X and β denotes the regression parameter, then a marginal model can be constructed based on the following three assumptions (Diggle et al., 2002):

- $E[Y_{ij} \mid X_{ij}] = \mu_{ij}$ depends on X_{ij} by $g(\mu_{ij}) = f(X_{ij}, \beta)$; Where $g(.)$ and $f(.)$ are known link and mean structure functions respectively
- $Var[Y_{ij} \mid X_{ij}] = \phi v(\mu_{ij})$; Where ϕ is a scale parameter and $v(\mu_{ij})$ is a

variance function

- $Corr[Y_{ij}, Y_{ik}] = \rho(\mu_{ij}, \mu_{ik}, \theta)$; Where $\rho(.)$ is a known function of marginal means at two different levels of X and a possible additional parameter θ

Depending on the mean structure-function $f(.)$ marginal models can be formulated as generalized linear models (GLM) or generalized nonlinear models (GNLM). The most advanced and popular approach to model longitudinal data marginally is through generalized estimating equations (GEE). One major convenience of using the marginal models is that parameter estimation is relatively less demanding compared to their counterparts in conditional models (Overall and Tonidandel, 2004; Zeger and Liang, 1986). We need to be careful about the fact that GEE treats correlation structure merely as a nuisance characteristic (Vonesh, 1992). Marginal models do not distinguish between within and between-subject variability in the response (Li, 2017). Results from marginal models cannot be interpreted at individual levels (Fitzmaurice and Ravichandran, 2008).

Random-effects models or mixed-effects models can be considered to be conditional models. This type of model attempts to capture the natural heterogeneity in the subjects arising from unobserved (latent) variables (Diggle et al., 2002). These models assume the response variable to be a function of explanatory variables with regression parameters that may vary by individuals. The response is assumed to be independent conditional on the subject-specific parameters (random effects). Distribution for the random effects is also specified in this modeling approach. Linear mixed-effects models (LMM), nonlinear mixed-effects models (NLMM), generalized linear mixed-effects models (GLMM), and generalized nonlinear mixed-effects models (GNLMM) are different possible forms of the random-effects models. Mixed-effects models enable researchers to interpret the results both at the individual level (through random effect

parameters) and at the population level (through fixed effect parameters). In general, this family of models is applicable to the incomplete and unbalanced dataset, and it is possible to derive marginal models from conditional random effect models (Lee et al., 2004). Two common criticisms of models like GLMM are their reliance on additional strict assumptions on random effects distribution and the computational intensity requirement to complete the fitting procedure (Muff et al., 2016).

Another variety of models for longitudinal data analysis is called transition (Markov) models (Diggle et al., 2002). These models assume the response variable to be a function of explanatory variables and the immediate prior values of the response. It can be considered to be a special case of marginal modes with prior outcomes as an additional explicit predictor of the response.

2.2.2 Mean function for the response

The selection of a mean function for the response is a vital component of longitudinal data analysis. Different forms of linear functions have been adopted to perform this task for a long time now. Perhaps it is due to the ease of computational procedures and the simplicity of interpretation of a linear-in-parameter mean function model (Wu, 2009). However, linear models only provide a local approximation to the true relationship between the response and covariates, when such a true relationship exists. On the other hand, nonlinear models sometimes called mechanistic models, provide a detailed understanding of the data generation process. Nonlinear models may offer better extrapolation of the responses since these are usually based on data-generating mechanisms. The estimated parameters in nonlinear models often have natural interpretations. Nonlinear models perform well in many settings like a biological system, population growth, or disease outbreak description. With the availability of modern

computing facilities, researchers should not restrict their focus to linear models only (Wu, 2009). Typically the parameters from nonlinear models are estimated with iterative algorithms. Davidian and Giltinan described a two-step method for statistical inferences from nonlinear models (Davidian and Giltinan, 2003). But this particular method requires a large number of observations per subject. In general, a nonlinear modeling approach should be avoided in cases where a reliable nonlinear function is not available.

2.2.3 Covariance structure

Finding an appropriate covariance structure for the models plays a key role in longitudinal data analysis. Incorrect estimates of the sampling variability and misleading scientific inferences can be made as a result of failure to properly capture the covariance among repeated measures (Fitzmaurice et al., 2012). A number of covariance structures are available for longitudinal response variables each having their own set of strengths and limitations (Fitzmaurice and Ravichandran, 2008).

An unstructured covariance structure has the following matrix form for the response Y_i,

$$Cov\left(Y_i\right) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1T} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2T} \\ \dots & \dots & \dots & \dots \\ \sigma_{T1} & \sigma_{T2} & \dots & \sigma_T^2 \end{bmatrix} \tag{2.1}$$

This structure does not make any assumptions about the variance and covariance. The major limitation of using this structure is that the number of parameters to be estimated grows rapidly with the number of occasions of mea-

surements. The use of this structure is also problematic when the dataset has mistimed measurements.

With the correlation structure, $Corr\,(Y_{ij}, Y_{ik}) = \rho,\ \rho \geq 0$; a compound symmetry covariance structure based on constant variance assumption has the following form,

$$Cov\,(Y_i) = \sigma^2 \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \dots & \dots & \dots & \dots \\ \rho & \rho & \dots & 1 \end{bmatrix} \tag{2.2}$$

Usually, the correlations are expected to decay with increasing separation in time; therefore making this constraint on the correlation among repeated measurements unappealing, and the constant variance assumption may not be valid in many settings.

An autoregressive (AR) covariance structure can be defined as follows:

$$Cov\,(Y_i) = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{T-1} \\ \rho & 1 & \rho & \dots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{T-3} \\ \dots & \dots & \dots & \dots & \dots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \dots & 1 \end{bmatrix} \tag{2.3}$$

Where the correlation, $Corr\,(Y_{ij}, Y_{ij+k}) = \rho^k,\ \rho \geq 0$.

This correlation structure is also known as discrete AR (1).

It can be considered to be a parsimonious covariance structure requiring only two parameters to be estimated for any number of measurement occasions. The autoregressive structure is appropriate in situations when the measurements are made at equal (or approximately equal) intervals of time. An-

other formulation of an autoregressive covariance structure, called exponential or continuous-time AR (1) structure (Fitzmaurice et al., 2012), can be used to deal with sparse datasets containing mistimed measurements. Assuming $Corr\left(Y_{ij}, Y_{ik}\right) = \rho^{|t_{ij}-t_{ik}|},\ \rho \geq 0$ and a constant variance σ^2 across all time-points, an exponential structure can be defined as

$$Cov\left(Y_i\right) = \sigma^2\rho^{|t_{ij}-t_{ik}|} = \sigma^2 exp\left(-\theta\left|t_{ij}-t_{ik}\right|\right); with\ \theta = -\log(\rho) \tag{2.4}$$

Fitzmaurice et al gave a detailed description of the different covariance structures for longitudinal data analysis in addition to the formations discussed above (Fitzmaurice et al., 2012).

2.2.4 Generalized nonlinear mixed models

A generalized nonlinear mixed modeling framework makes it possible to fit a wide range of nonlinear models by a relatively fast and robust method (Lane, 1996). This framework combines the generalized nonlinear models with random effects models. The resulting generalized nonlinear mixed model (GNLMM) is considered to be the nonlinear version of the commonly used model in longitudinal studies, the generalized linear mixed model (GLMM).

The GNLMMs can accommodate both continuous and discrete responses in the model. The effects of unobserved covariates on the response can be taken into account in GNLMM, thus allowing for individual-level interpretation of the results.

Let, $Y_{iT_i} = [Y_{i1}, Y_{i2}, \ldots, Y_{iT_i}]'_{T_iX1}$ be a vector of responses for the i-th subject (i=1,2,..,I),

X_{iT_i} and Z_{iT_i} represent the vectors of fixed effects and random effects respectively,

β and b_i represent the vectors of fixed effects and random effects parameter respectively.

Now assuming the independence of Y_{iT_i} conditional on the random effects and fixed effects, a GNLMM can be written as

$$g\left(\mu_{iT_i}\right) = f(X_{ij}, \beta, b_i) \tag{2.5}$$

where $\mu_{iT_i} = E\left[Y_{iT_i}|X_i, Z_i, b_i\right]$ is the conditional mean of Y_{iT_i}. $g\left(.\right)$ and $f(.)$ are known link and mean functions respectively.

2.2.5 Tweedie exponential family of distributions

An exponential family is a parametric set of probability distributions of a certain mathematically convenient form (Clark and Thayer, 2004). If X is a random variable from exponential family and θ is the parameter space to define its distribution, then the form of the distribution can be written as

$$f\left(x|\theta\right) = h(x)\exp[\eta\left(\theta\right)T\left(x\right) - A(\theta)] \tag{2.6}$$

where $h(x)$, $\eta(\theta)$, $T(x)$ and $A(\theta)$ are known functions.

The Tweedie family of distributions is named after a British statistician, Maurice Tweedie by another statistician from Denmark, Bent Jørgensen (Jørgensen, 1987). It is a subset of the exponential family. One of the distinct features of this family of distributions is that it can have a cluster of data points at zero (called a "point mass"), which is particularly useful for modeling claims in the insurance industry, in medical/genomic testing, rainfall forecasting, or anywhere else there is a mixture of zeros and non-negative data points. This family also contains a wide range of continuous and discrete distributions.

If Y is a random variable having the following distributional form:

$$f\left(y,\mu,\tau^2\right)=\begin{cases} c_p(y,\tau^2)exp[\frac{1}{\tau^2}(\frac{y\mu^{1-p}}{1-p}-\frac{\mu^{2-p}}{2-p})] & \text{if } p\neq 1,2 \\ c_p(y,\tau^2)exp[-\frac{1}{\tau^2}(\frac{y}{\mu}+\log(\mu)] & \text{if } p=2 \\ c_p(y,\tau^2)exp[(y\log(\mu)-\mu] & \text{if } p=1 \end{cases} \tag{2.7}$$

Then, we can say that Y follows a Tweedie exponential distribution $Tw_p\left(\mu,\tau^2\right)$ with mean, $E\left[Y\right]=\mu$ and variance, $Var\left[Y\right]=\tau^2\mu^p$ (Ma and Jørgensen, 2007). This family of distributions is also known as the power-variance family because of the unique relationship between mean and variance depending on the power, p. Many familiar distributions are special cases of the Tweedie distribution, which can be resulted based on the value of p. Some of these scenarios are given in table 2.2:

Table 2.2: **Tweedie distributions and power parameter**

Power of mean in variance	**Corresponding distribution**
p = 0	Normal distribution
p = 1	Poisson distribution
$1 < p < 2$	Compound Poisson
p = 2	Gamma distribution
$2 < p < 3$	Positive stable distributions
p = 3	Inverse Gaussian distribution

Several earlier studies have included detailed descriptions of Tweedie distributions (Jørgensen, 1987; Ma and Jørgensen, 2007; Jorgensen, 1997; Tweedie, 1984).

2.2.6 Joint models for longitudinal data

In longitudinal studies, measurements are often collected for more than one outcome for each subject. One approach to analyzing such datasets is to model these outcomes separately in different models. But following this approach

cannot help if the interest lies in an association structure or the joint test for fixed effects (Verbeke and Davidian, 2009). Jointly modeling multiple responses is a solution to overcome this obstacle. There are several methods to construct joint models for multiple responses.

Let $Y1$ and $Y2$ be two responses measured on all the subjects under a longitudinal study.

It may be possible to formulate joint models by specifying a joint density, $f(y_1, y_2)$ directly. It will allow for direct inferences for the marginal characteristics of $Y1$, $Y2$, and their associations. But it is difficult to extend to higher dimensions or if $Y1$ and $Y2$ represent two different types (For example, discrete and continuous) of outcomes.

Conditional joint models can be developed by factorizing $f(y_1, y_2)$ as

$$f\left(y_1, y_2\right) = f\left(y_1|y_2\right) f\left(y_2\right) = f\left(y_2|y_1\right) f\left(y_1\right) \tag{2.8}$$

This will reduce the modeling tasks to specifying models for each of the outcomes separately. No direct marginal inferences are possible from the resulting conditional model. Moreover, effects on $Y1$ may be affected as a result of conditioning on $Y2$.

Shared parameter models or random-effects models are another widely used joint modeling approach for longitudinal outcomes. Assuming that latent variables are shared by $Y1$ and $Y2$ this type of model can be formulated. Let b denotes a vector of random effects, with density $f(b)$ and there exists conditional independence between the outcomes given b. Then the joint density can be defined in the form

$$f\left(y_1,\ y_2\right) = f\left(y_1,\ y_2|b\right) db = f(y_1|b) f(y_2|b) f(b)\ db \tag{2.9}$$

In this method of joint modeling $Y1$ and $Y2$ can represent two different types of responses and extension to higher dimensions is straightforward. These models rely on a very strong assumption about the association between $Y1$ and $Y2$, which is problematic in some cases in higher dimensions. In order to overcome this difficulty (Verbeke and Davidian, 2009) suggested assuming separate but correlated latent variables for $Y1$ and $Y2$. But it comes with the high cost of computational intensity as it will increase the dimensionality of b with the number of outcomes being modeled. Shared parameter models are not flexible in the sense that $f(b)$ is restricted to be normal distribution in most cases (Verbeke and Davidian, 2009). It is desirable to find more flexible shared random-effects models due to this restriction.

Some methods based on dimension reduction techniques are also available (Verbeke and Davidian, 2009). In such methods, in the initial stage the dimensionality of the response vector is reduced by using factor analysis or principal component methods. Finally, the resulting principal factors are analyzed using any of the classical (longitudinal) models. The results of this type of analysis will only allow inferences about principal factors, not about original outcome variables. In addition, strong constraints, such as observations taken at arbitrary time points and unequal numbers of measurements for different subjects, are needed in cases of highly unbalanced longitudinal data.

This thesis focuses on deriving joint models for the responses from the Tweedie exponential family based on shared random effects with flexible distributional assumptions about random effects. Li proposed linear joint models of this type applied to a balanced longitudinal dataset (Li, 2017). Snow developed an univariate version of these models (Snow, 2019). Ma et al presented Tweedie generalized linear models allowing a wide range of skewness and covariance structures for different data types (Ma et al., 2018). This thesis introduces

nonlinear joint models applicable for longitudinal datasets.

Chapter 3

Nonlinear Joint Model for Longitudinal Data

This chapter introduces the joint mixed-effect models for more than one response from a broad family of exponential distributions known as Tweedie. A number of familiar distributions including Normal, Poisson, Compound Poisson, Gamma, and Inverse Gaussian are special cases of the Tweedie distribution. The proposed model accommodates both fixed and random effects to allow for interpretation at population and individual levels respectively. The models are based on three assumptions on random effects and responses. The following sections in this chapter will describe the model components in detail.

3.1 Model statement

Let Y_{ijt} stands for the i-th (i=1,2,. . . .,I) individuals (cluster) measurement on j-th (j=1,2,. . . .,J) longitudinal response at t-th timepoint (sub-cluster) (t=1,2,. . . .,T_i). The vector of responses, Y_{ijt} can be represented with, $Y = (Y_{111}, Y_{112}, \ldots, Y_{11T_1},$ $Y_{121}, Y_{122}, \ldots, Y_{12T_1}, \ldots, Y_{1J1}, Y_{1J2}, \ldots Y_{1JT_1}, \ldots., Y_{IJ1}, \ldots., Y_{IJT_i})'$.

Two different random effects, subject-specific U_i and time-specific, V_{ijt} are

considered in this model.

The three key assumptions based on which the model is constructed are given below:

Assumption 1: This assumption deals with subject-specific random effects $U_1, U_2, \ldots, U_I$. It is assumed that $U_{i'}$s are positive, independently and identically distributed with mean, $E[U_i] = 1$ and variance, $Var[U_i] = \sigma^2$

Assumption 2: This assumption is about time-specific random effects, $V_{111}, V_{112}, \ldots, V_{11T_1}, V_{121}, V_{122}, \ldots, V_{12T_1}, \ldots, V_{1J1}, V_{1J2}, \ldots, V_{1JT_1}, \ldots, V_{IJ1}, \ldots, V_{IJT_i}$. It is assumed that $V_{ijt'}$'s are positive and have conditional moment structures depending on subject-specific random effect U as follows,

$$E[V_{ijt}|U] = U$$

$$\text{and } Cov[V_{ijt}, V_{i'j't'}|U] = \begin{cases} \tau_j^2 \rho_j(t,t')U_i, \text{if } i = i' \text{and} j = j' \\ 0, \text{Otherwise} \end{cases}$$

Where the correlation structure among the repeated measures of j-th response, $\rho_{j(t,t')}$ can be decided depending on the type of longitudinal data. For example, if we are dealing with an unbalanced longitudinal dataset where subjects of interest are measured at different number of timepoints and irregular intervals, we may consider Complete independence or Continuous autoregressive with order 1 (CAR (1)) correlation structure.

Thus, measurement at a particular timepoint on a subject is considered to be a sub-cluster within a cluster (subject). On the other hand, to deal with a balanced longitudinal dataset we can pick a correlation structure from a number of available structures including Exchangeable (or Compound symmetry), Autoregressive with order 1 (AR (1)) and Unstructured.

Assumption 3: This assumption is about the response variables in the model. The responses are assumed to be independent conditional on the random effects U and V. The conditional distribution of the response is assumed to follow Tweedie distribution. The resulting distribution can be written as $Y_{ijt}|(U,V) \sim Tw_{p_j}(\mu_{ijt}V_{ijt}, \varepsilon_j^2 V_{ijt}^{1-p_j})$. where $Tw_{p_j}(\mu_{ijt}V_{ijt}, \varepsilon_j^2 V_{ijt}^{1-p_j})$ denotes the Tweedie distribution with index parameter p, expected value $\mu_{ijt}V_{ijt}$ and dispersion parameter $\varepsilon_j^2 V_{ijt}^{1-p_j}$. This set of exponential distributions can also be termed as the power variance family with,

$$\text{mean, } E\left[Y_{ijt}\,|(U,V)\right] = \mu_{ijt}V_{ijt}$$

and variance, $Var\left[Y_{ijt}\,|(U,V)\right] = \text{Dispersion Parameter} * (\text{Expected Value})^{p_j}$

$$= \varepsilon_j^2 V_{ijt}^{1-p_j} * (\mu_{ijt}V_{ijt})^{p_j}$$

$$= \varepsilon_j^2 V_{ijt}\mu_{ijt}^{p_j}$$

Where, $\mu_{ijt} = f(x_{ijt}, \beta)$ is a nonlinear function of covariates x_{ijt} and regression parameters β. The purpose of this function is to capture the mean structure of the responses in the model.

3.2 Moment structures

The estimators and the predictors of the random effects introduced in the following sections are derived based on the marginal moment structures. The following two propositions in probability (Weiss, 2005a; Ross et al., 2006) theory are implemented to find the moment structures,

Proposition 1: If X is a random variable having defined expected value $E[X]$

and Y is any random variable on the same probability space, then

$$E[X] = E[E[X|Y]] \tag{3.1}$$

The law of iterated expectations is one of the names among others frequently used to present this proposition.

Proposition 2: If X, Y, and Z are random variables on the same probability space, and X and Y have finite covariance, then,

$$Cov[X, Y] = E[Cov[X, Y|Z]] + Cov[E[X|Z], E[Y|Z]] \tag{3.2}$$

It is frequently termed the law of total covariance or conditional covariance formula in probability literatures.

The moment structures are calculated using (3.1 and 3.2), the conditional moments of the responses and time-specific random effects, and unconditional moments of subject-specific random effects in the defined model. In the rest of the parts of this chapter, Kronecker delta $\delta_{(i,i')}$ is used in different equations, which equals 1 when $i = i'$ and 0 when $i \neq i'$.

3.2.1 Marginal moments of time-specific random effects

The expectation of the time-specific random effects, V_{ijt} can be expressed as

$$\begin{aligned} E[V_{ijt}] &= E[E[V_{ijt}|U]] \\ &= E[U_i] \\ &= 1 \end{aligned} \tag{3.3}$$

The covariance of the time-specific random effects, V_{ijt} can be expressed as

$$
\begin{aligned}
Cov\left[V_{ijt}, V_{i'j't'}\right] &= E\left[Cov\left[V_{ijt}, V_{i'j't'}|U\right]\right] + Cov\left[E\left[V_{ijt}|U\right], E\left[V_{i'j't'}|U\right]\right] \\
&= \left[\delta_{(i,i')}\delta_{(j,j')}\tau_j\rho_{j(t,t')}U_i\right] + \delta_{ii'}Cov\left[U_i, U_{i'}\right] \\
&= \delta_{(i,i')}\left[\delta_{(j,j')}\tau_j^2\rho_{j(t,t')} + \sigma^2\right]
\end{aligned} \tag{3.4}
$$

Four possible cases based on 3.4 can be calculated as follows,

Case 1: If $i = i'$, $j = j'$ and $t = t'$,

$$
\begin{aligned}
Cov\left[V_{ijt}, V_{ijt}\right] &= Var[V_{ijt}] \\
&= \tau_j^2\ \rho_j(t,t)\ + \sigma^2 \\
&= \tau_j^2\ \times\ 1 + \sigma^2 \\
&= \tau_j^2\ + \sigma^2
\end{aligned}
$$

Case 2: If $i = i'$, $j = j'$ and $t \neq t'$,

$$
Cov\left[V_{ijt}, V_{ijt'}\right] = \tau_j^2\ \rho_j(t,t') + \sigma^2
$$

Case 3: If $i = i'$, $j \neq j'$ and $t \neq t'$,

$$
\begin{aligned}
Cov\left[V_{ijt}, V_{ij't'}\right] &= 0\ \times\ \tau_j^2\ \rho_j(t,t') + \sigma^2 \\
&= \sigma^2
\end{aligned}
$$

Case 4: In all other cases,

$$
Cov\left[V_{ijt}, V_{i'j't'}\right] = 0
$$

3.2.2 Marginal moments of response variables

The expectation of the responses, Y_{ijt} can be expressed as

$$\begin{aligned} E\left[Y_{ijt}\right] &= E\left[E\left[Y_{ijt}|(U,V)\right]\right] \\ &= \mu_{ijt} E\left[V_{ijt}\right] \\ &= \mu_{ijt} \end{aligned} \tag{3.5}$$

The covariance of the responses, Y_{ijt} can be expressed as

$$\begin{aligned} Cov\left[Y_{ijt}, Y_{i'j't'}\right] &= E\left[Cov\left[Y_{ijt}, Y_{i'j't'}|(U,V)\right]\right] + Cov\left[E\left[Y_{ijt}|(U,V)\right], E\left[Y_{i'j't'}|(U,V)\right]\right] \\ &= \delta_{(i,i')}\delta_{(j,j')}\delta_{(t,t')}\varepsilon_j^2\mu_{ijt}^{p_j} + \mu_{ijt}\mu_{i'j't'}\delta_{(i,i')}\left[\delta_{(j,j')}\tau_j^2\rho_{j(t,t')} + \sigma^2\right] \end{aligned} \tag{3.6}$$

Four possible cases based on 3.6 can be calculated as follows,

Case 1: If $i = i'$, $j = j'$ and $t = t'$,

$$\begin{aligned} Cov\left[Y_{ijt}, Y_{ijt}\right] &= Var[Y_{ijt}] \\ &= \varepsilon_j^2\ \mu_{ijt}^{p_j} + \mu_{ijt}^2(\sigma^2 + \tau_j^2) \end{aligned}$$

Case 2: If $i = i'$, $j = j'$ and $t \neq t'$,

$$Cov\left[Y_{ijt}, Y_{ijt'}\right] = 0\ \times\ \varepsilon_j^2\ \mu_{ijt}^{p_j} + \mu_{ijt}\ \mu_{ijt'}(\sigma^2 + \tau_j^2\rho_j(t,t'))$$

Case 3: If $i = i'$, $j \neq j'$ and $t \neq t'$,

$$\begin{aligned} Cov\left[Y_{ijt}, Y_{ij't'}\right] &= 0\ \times\ \varepsilon_j^2\ \mu_{ijt}^{p_j} + \mu_{ijt}\ \mu_{ij't'}\sigma^2 + 0\ \times\tau_j^2\rho_j(t,t')) \\ &= \mu_{ijt}\ \mu_{ij't'}\sigma^2 \end{aligned}$$

Case 4: In all other cases,

$$Cov\,[Y_{ijt}, Y_{i'j't'}\,] = 0$$

3.3 The best linear unbiased predictor of the random effects

This section adopts the orthodox Best Linear Unbiased Predictor (BLUP) for the subject-specific and time-specific random effects presented in the previous sections. The term "Orthodox" is used to distinguish it from the modal predictors of the random effects (Ma, 1999).

3.3.1 Subject-specific random effects

The BLUP of $U_1, U_2, \ldots, U_I$ given the response, Y can be expressed as

$$\hat{U} = E\,[U] + Cov\,[U,\ Y\,]\,Var^{-1}\,[Y]\,[Y - E\,[Y]] \tag{3.7}$$

where

$E[U]$ is a unit column vector having IJ elements,

$Cov[U,\ Y\ ]$ is a matrix containing the marginal covariance between U and Y,

The marginal covariance between subject-specific random effects, U_i and re-

sponse variables, Y_{ijt} can be derived as

$$\begin{aligned} Cov\left[U_{i'}, Y_{ijt}\right] = & \; E\left[Cov\left[U_{i'}, Y_{ijt}|(U,V)\right]\right] + Cov\left[E\left[U_{i'}|(U,V)\right], E\left[Y_{ijt}|(U,V)\right]\right] \\ = & \; \delta_{(i,i')}\delta_{(j,j')}Cov\left[U_{i'}, \mu_{ijt}V_{ijt}\right] \\ = & \; \mu_{ijt}\delta_{(i,i')}\delta_{(j,j')}Cov\left[U_{i'}, V_{ijt}\right] \\ = & \; \mu_{ijt}\delta_{(i,i')}\delta_{(j,j')}\left[E\left[Cov\left[U_{i'}, V_{ijt}|U\right]\right] + Cov\left[E\left[U_{i'}|U\right], E\left[V_{ijt}|U\right]\right]\right] \\ = & \; \mu_{ijt}\delta_{(i,i')}\left[Cov\left[U_{i'}, U_i\right]\right] \\ = & \; \mu_{ijt}\delta_{(i,i')}\sigma^2 \end{aligned} \tag{3.8}$$

$Var^{-1}[Y]$ is the inverse of the marginal variance-covariance matrix of Y,
$[Y - E[Y]]$ is a column vector of length IJ with the differences between the measurements on the responses and their respective means.

3.3.2 Time-specific random effects

Similarly an expression for the BLUP of the time-specific random effects $V_{111}, V_{112}, \ldots, V_{11T_1}, V_{121}, V_{122}, \ldots, V_{12T_1}, \ldots, V_{1J1}, V_{1J2}, \ldots, V_{1JT_1}, \ldots, V_{IJ1}, \ldots, V_{IJT_i}$ can be written as

$$\hat{V} = \; E[V] + \; Cov[V, \; Y \;] \, Var^{-1}[Y] \, [Y - E[Y]] \tag{3.9}$$

For the joint models defined earlier,
$E[V]$ is a unit column vector having IJ elements,
$Cov[V, \; Y \;]$ is a matrix containing the marginal covariance between U and Y,
The marginal covariance between time-specific random effects, V_{ijt} and re-

sponse variables, Y_{ijt} can be expressed as

$$\begin{aligned} Cov\left[V_{i'j't'}, Y_{ijt}\right] &= E\left[Cov\left[V_{i'j't'}, Y_{ijt}|(U,V)\right]\right] + Cov\left[E\left[V_{i'j't'}|(U,V)\right], E\left[Y_{ijt}|(U,V)\right]\right] \\ &= \delta_{(i,i')}\delta_{(j,j')}Cov\left[V_{i'j't'}, \mu_{ijt}V_{ijt}\right] \\ &= \mu_{ijt}\delta_{(i,i')}\delta_{(j,j')}Cov\left[V_{i'j't'}, V_{ijt}\right] \\ &= \mu_{ijt}\delta_{(i,i')}Cov\left[\delta_{(j,j')}\delta_{(j,j')}\tau_j^2\rho_{j(t,t')} + \sigma^2\right] \end{aligned} \tag{3.10}$$

$Var^{-1}[Y]$ and $[Y - E[Y]]$ carry the same meaning as it is in the previous subsection.

A convenient version of matrix expression for the matrices introduced in this section is available in a previous study (Li, 2017).

3.4 Parameter Estimation

In this section, we discuss the estimation methods for regression parameters, dispersion parameters, and correlation parameters presented in the Nonlinear Joint Model. The Newton scoring function is used to find estimators for the regression parameters. The estimators for dispersion parameters and correlation parameters are derived with the adjusted Pearson estimator (Ma, 1999).

3.4.1 Regression parameters

In order to find the regression parameters β, a partially observed joint log-likelihood function of the defined model for the data and random effects is derived. This function is differentiated with respect to β to find the partially observed joint score function (Ma and Jørgensen, 2007). An unbiased estimating function, $\psi(\beta)$ for the regression parameters is achieved by replacing the

random effects with their BLUPs in the score function as follows:

$$\psi(\beta) = \sum_{i=0}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T_i} X'_{ijt}\frac{\mu_{ijt}^{1-p_j}(\beta)}{\varepsilon_j^2}\left[Y_{ijt} - \hat{V}_{ijt}(\beta)\mu_{ijt}(\beta)\right] \tag{3.11}$$

The solution to the equation, $\psi(\beta) = 0$ gives the estimates of the regression parameters. The resulting estimates are consistent and asymptotically normal under mild conditions (Ma, 1999). The distribution of the estimated parameter is normal (Ma and Jørgensen, 2007) having asymptotic mean β and asymptotic variance with the negative inverse of the sensitivity matrix,

$$S(\beta) = E_\beta\left[\frac{\partial\psi(\beta)}{\partial\beta}\right]$$

If X is the model matrix, then the explicit form of the estimated score function and the sensitivity matrix (Ma and Jørgensen, 2007) can be written as

$$\psi(\beta) = -X'diag[E[Y]]Var - 1[Y][Y - E[Y]] \tag{3.12}$$

$$S(\beta) = -X'diag\left[E\left[Y\right]\right]Var^{-1}\left[Y\right]diag\left[E\left[Y\right]\right]X \tag{3.13}$$

In 3.12, $\psi(\beta)$ is a vector having length equal to the number of parameters to be estimated from the joint model of interest and 3.13 gives a square matrix, $S(\beta)$ having order as the length of $\psi(\beta)$.

The estimates of the parameters as a solution to $\psi(\beta) = 0$ can be found by using the iterative newton scoring algorithm with appropriate initial values as follows

$$\beta_{k+1} = \beta_k - S^{-1}(\beta_k)\psi(\beta_k) \tag{3.14}$$

3.4.2 Dispersion parameters

The dispersion parameters are estimated with two components. First, the method of moments is used to estimate the random effects. Then, a bias correction is added to account for the use of BLUP estimates of random effects instead of their true value. An iterative equation (Li, 2017) which estimates the variance of subject-specific random effects σ^2, is

$$\sigma_k^2 = \frac{1}{I}\sum_{i=1}^{I}\left[\left(\hat{U}_i - 1\right)^2 + \hat{\sigma}_{k-1}^2 - \hat{\sigma}_{k-1}^4 \mu_i' Var^{-1}(Y)\mu_i\right] \tag{3.15}$$

Similarly, estimates for the parameters τ_j^2 and ε_j^2 from the dispersion parameters of the conditional time-specific random effects and conditional responses respectively are given below,

$$\begin{aligned}\tau_{j,k}^2 = & \frac{1}{IT_i}\sum_{i=1}^{I}\sum_{t=1}^{T_i}\left[\left(\hat{V}_{ijt} - \hat{U}_i\right)^2 + \tau_{j,k-1}^2\right] \\ & - \frac{1}{IT_i}\sum_{i=1}^{I}\sum_{t=1}^{T_i}\left[\sigma^4 \mu_{ij}' Var^{-1}\left(Y_{ij}\right)\left(\sigma^2\mu_{ij}\right) cov\left(V_{ijt}, Y_{ij}\right) Var^{-1}\left(Y_{ij}\right) cov\left(Y_{ij}, V_{ijt}\right)\right] \\ & + \frac{1}{IT_i}\sum_{i=1}^{I}\sum_{t=1}^{T_i}\left[2cov\left(U_i, Y_{ij}\right) Var^{-1}\left(Y_{ij}\right) cov\left(Y_{ij}, V_{ijt}\right)\right]\end{aligned} \tag{3.16}$$

$$\begin{aligned}\varepsilon_{j,k}^2 = & \frac{1}{IT_i}\sum_{i=1}^{I}\sum_{t=1}^{T_i} \\ & \frac{1}{\mu_{ijt}^{p_j}}\left[\left(y_{ijt} - \mu_{ijt}\hat{V}_{ijt}\right)^2 + \mu_{ijt}^2\left(\sigma^2 + \tau_j^2\right) - \mu_{ijt}^2 cov\left(V_{ijt}, Y_{ij}\right) Var^{-1}\left(Y_{ij}\right) cov\left(Y_{ij}, V_{ijt}\right)\right]\end{aligned} \tag{3.17}$$

3.4.3 Correlation parameters

If each of the subjects in a longitudinal study have an equal number of observations T (balanced longitudinal dataset), a basic layout for a correlation matrix can be organized as

$$\begin{aligned} R_j &= \left[\rho_{j(t,t')}\right]_{TxT} \\ &= \begin{bmatrix} 1 & \rho_{j(1,2)} & \rho_{j(1,3)} & \cdots & \rho_{j(1,T-1)} \\ \rho_{j(2,1)} & 1 & \rho_{j(2,3)} & \cdots & \rho_{j(2,T-1)} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_{j(T-1,1)} & \rho_{j(T-1,2)} & \rho_{j(T-1,3)} & \cdots & 1 \end{bmatrix} \end{aligned} \tag{3.18}$$

The estimators of the correlation parameters are derived in the same manner as the dispersion parameters. Depending on the dataset, an appropriate correlation structure can be chosen from the usual structures such as Exchangeable, Unstructured, Complete Independence and AR (1). Li provides the derivation of explicit formulae to calculate $\rho_{j(t,t')}$ from different correlation structures (Li, 2017). In the case of the unstructured correlation structure,

$$\begin{aligned} \rho_{j(t,t')} &= \frac{Cov\left[\left(V_{ijt}-U_i\right)\left(V_{ijt'}-U_i\right)\right]}{\left[Var\left(V_{ijt}-U_i\right)Var\left(V_{ijt'}-U_i\right)\right]^{1/2}} \\ &= \frac{Cov\left[\left(\hat{V}_{ijt}-\hat{U}_i\right)\left(\hat{V}_{ijt'}-\hat{U}_i\right)+b_{j,i(t,t')}\right]}{[\left\{Var\left(\hat{V}_{ijt}-\hat{U}_i\right)+b_{j,i(t,t)}\right\}Var\left(\hat{V}_{ijt'}-\hat{U}_i\right)+b_{j,i(t',t')}]^{1/2}} \end{aligned} \tag{3.19}$$

where, $b_{j,i(t',t')}$ is the correction term. The simplified version of this correction term can be written as

$$\rho_{j(t,t')}\tau_j^2 - [Cov\left(\hat{V}_{ijt},\hat{V}_{ijt'}\right)+\sigma^2\mu'_{ij}Var^{-1}\left(Y_{ij}\right)\left(\sigma^2\mu_{ij}\right)-cov\left(Y_{ij},V_{ijt}\right)-cov(Y_{ij},V_{ijt'})]$$

The left hand side in 3.19 can be estimated using adjusted Pearson estimator as follows:

$$\hat{\rho}_{j(t,t')} = \frac{\sum_{i=1}^{I}\left[\left(\hat{V}_{ijt}-\hat{U}_i\right)\left(\hat{V}_{ijt'}-\hat{U}_i\right)+b_{j,i(t,t')}\right]}{\left[\left\{\sum_{i=1}^{I}\left(\hat{V}_{ijt}-\hat{U}_i\right)^2+b_{j,i(t,t)}\right\}\left\{\sum_{i=1}^{I}\left(\hat{V}_{ijt'}-\hat{U}_i\right)^2+b_{j,i(t',t')}\right\}\right]^{1/2}} \tag{3.20}$$

Following (Li, 2017), the for AR (1) correlation structure, ρ_j can be estimated using the following formula:

$$\hat{\rho}_j = \frac{\sum_{i=1}^{I}\left[\left(\hat{V}_{ijt}-\hat{U}_i\right)\left(\hat{V}_{ij(t+1)}-\hat{U}_i\right)+b_{j,i(t,(t+1))}\right]}{\left[\sum_{i=1}^{I}\left(\hat{V}_{ijt}-\hat{U}_i\right)^2+b_{j,i(t,t)}\sum_{i=1}^{I}\left(\hat{V}_{ijt'}-\hat{U}_i\right)^2+b_{j,i((t+1),(t+1))}\right]^{1/2}} \tag{3.21}$$

In the case of discrete AR (1),$\rho_{j(t,t')} = \rho_j^{|t-t'|}, t \neq t'$; Where, t and t' represent the order of column and row respectively in the correlation matrix R_j.
In this thesis we used discrete AR (1) and Exchangeable correlation structures for data analysis.

Chapter 4

Data Analysis

To demonstrate our proposed modeling approach, it is applied to an anti-hypertensive drug dataset. The responses from test and reference drug formulations are modeled jointly. Four joint models based on the combinations of non-detectable response measure imputation and correlation structures are fitted to explore the pattern of drug concentration level changes in blood. One of the major objectives of this data analysis is to check the bioequivalence between the responses under consideration. The result from the univariate models corresponding to the two drug outcomes is presented along with the proposed joint model outputs for comparison purposes. The entire computational procedure is completed in the RStudio software cloud. An R package gnm is used to find the initial values of the nonlinear models required to start the Newton-Raphson algorithm. We used the Taylor expansion based linear approximation to the nonlinear functions to fit nonlinear regression models.

4.1 Losartan data

Losartan is a drug used to treat hypertension by blocking angiotensin II receptors (a class of G protein-coupled receptors). The use of this drug is also evident in the treatment of patients suffering from other diseases such as diabetic kidney disease, heart failure, and left ventricular enlargement. This dataset was

constructed from a crossover randomized study that administered two different formulations of the drug (test and reference) to 24 healthy volunteers (Willemsen et al., 2017). In order to ensure no carry-over effect, the drugs were given to each participant on different days in a randomized order. The administered drug concentration levels in blood were measured at 23 fixed timepoints including the baseline. The layout of the analysed dataset is presented in the following table,

Table 4.1: **Data structure**

VoI	**Concentration 1**	**Time 1**	**Concentration 2**	**Time 2**	**Order**
1	0	0	0	0	1
1	90.09	0.33	130.39	0.33	1
⋮	⋮	⋮	⋮	⋮	⋮
1	0	24	0	24	1
2	0	0	0	0	0
2	14.17	0.33	42.88	0.33	0
⋮	⋮	⋮	⋮	⋮	⋮
2	0	24	0	24	0
⋮	⋮	⋮	⋮	⋮	⋮
24	0	0	0	0	1
24	736.87	0.33	158.85	0.33	1
⋮	⋮	⋮	⋮	⋮	⋮
24	3.94	24	0	24	1

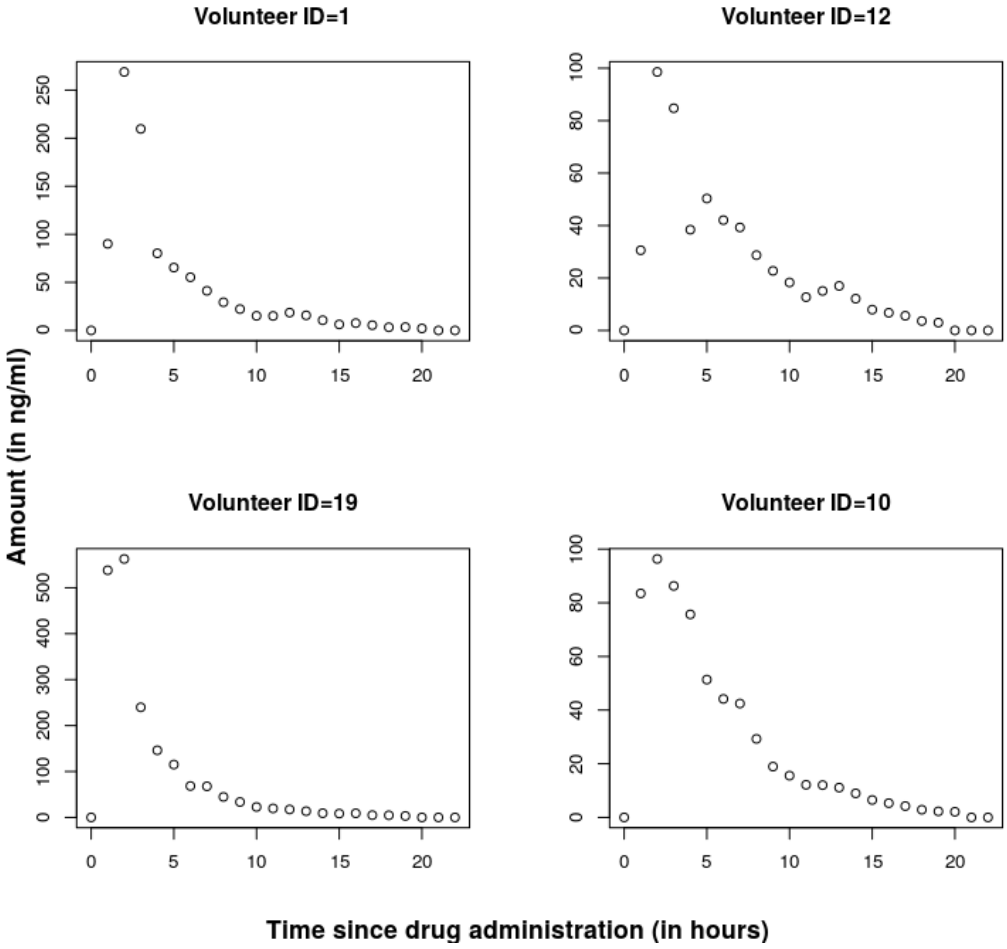
The heading VoI stands for volunteer ID and concentration levels in blood resulting from reference and test drug formulations are represented by the headings Concentration 1 and Concentration 2 respectively. Measurements were taken at 23 fixed timepoints: 0, 20 minutes, 30 minutes, 40 minutes, 60 minutes, 80 minutes, 100 minutes, 2hrs, 2.5hrs, 3hrs, 3.5hrs, 4hrs, 4.5hrs, 5hrs, 5.5hrs, 6hrs, 6.5hrs, 7hrs, 8hrs, 9hrs, 10hrs, 14hrs and 24hrs. The order variable presents the randomized order used in the crossover study. There are no

missing values in the dataset. The above table shows that we have a balanced longitudinal dataset with complete information on 24 subjects at 23 timepoints.

4.2 Exploratory data analysis

Figure 4.1 shows the change in reference drug concentration level for 4 randomly selected volunteers after getting the high blood pressure treatment. To depict the overall trend, the time plot in **Figure 4.2** combines all the data points.

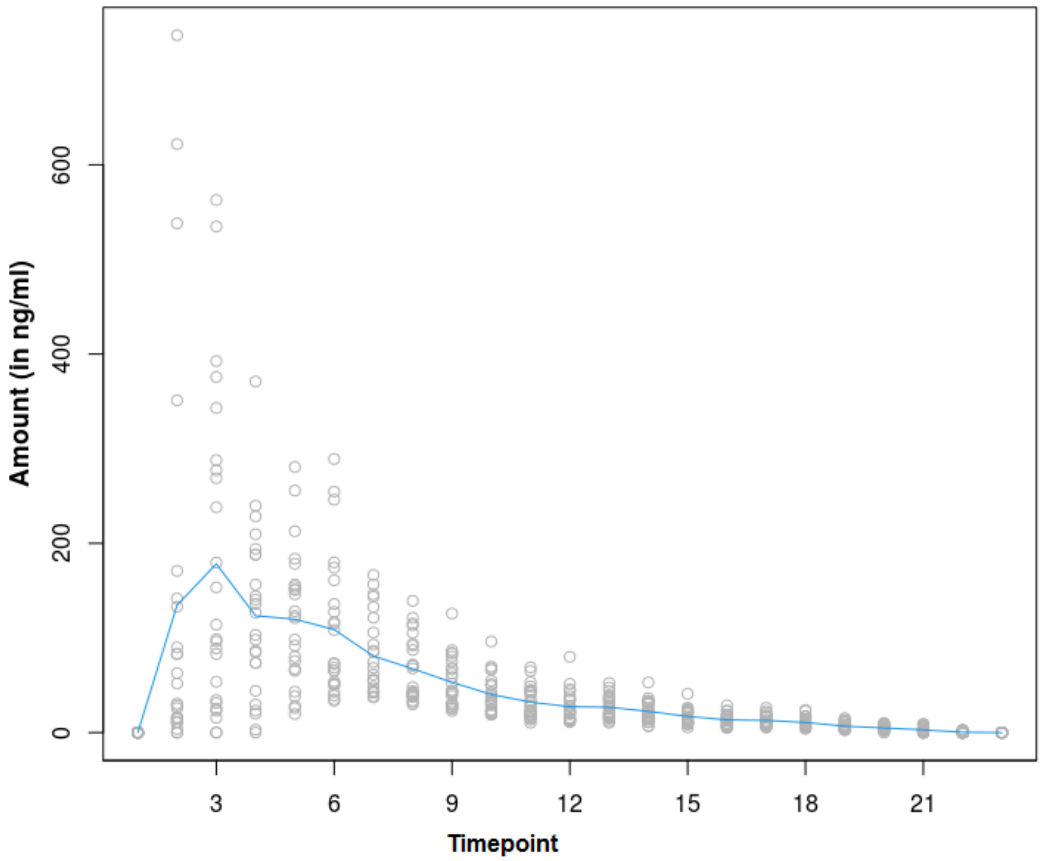
Figure 4.1: **Reference drug concentration level for four randomly selected volunteers**



The blue line represents the mean concentration level of the reference drug in

blood at different timepoints. It is evident from the graphs mentioned earlier that the concerned drug concentration level increases rapidly immediately after administering the drug into body, then it starts decreasing considerably after the third timepoint (0.50 hrs) for most of the subjects. Finally, it starts becoming stable after the ninth timepoint (2 hrs) for all the participants under study.

Figure 4.2: **Reference drug concentration level for all volunteers**



The test drug concentration level in blood for the randomly selected individuals is portrayed in **Figure 4.3**. The overall change in test drug concentration level for all the patients under study is displayed in **Figure 4.4**. This time plot indicates quite a similar pattern like the reference drug concentration level in blood for most of the subjects enrolled in the study. However, a few participants have shown totally different changing pattern in drug concentration level over

time compared to others. For example, the individual with ID=10 responded with symmetric changing pattern in test drug concentration. Considerable variation across individuals is also indicated by both figures presented below.

Figure 4.3: **Test drug concentration level for four randomly selected volunteers**

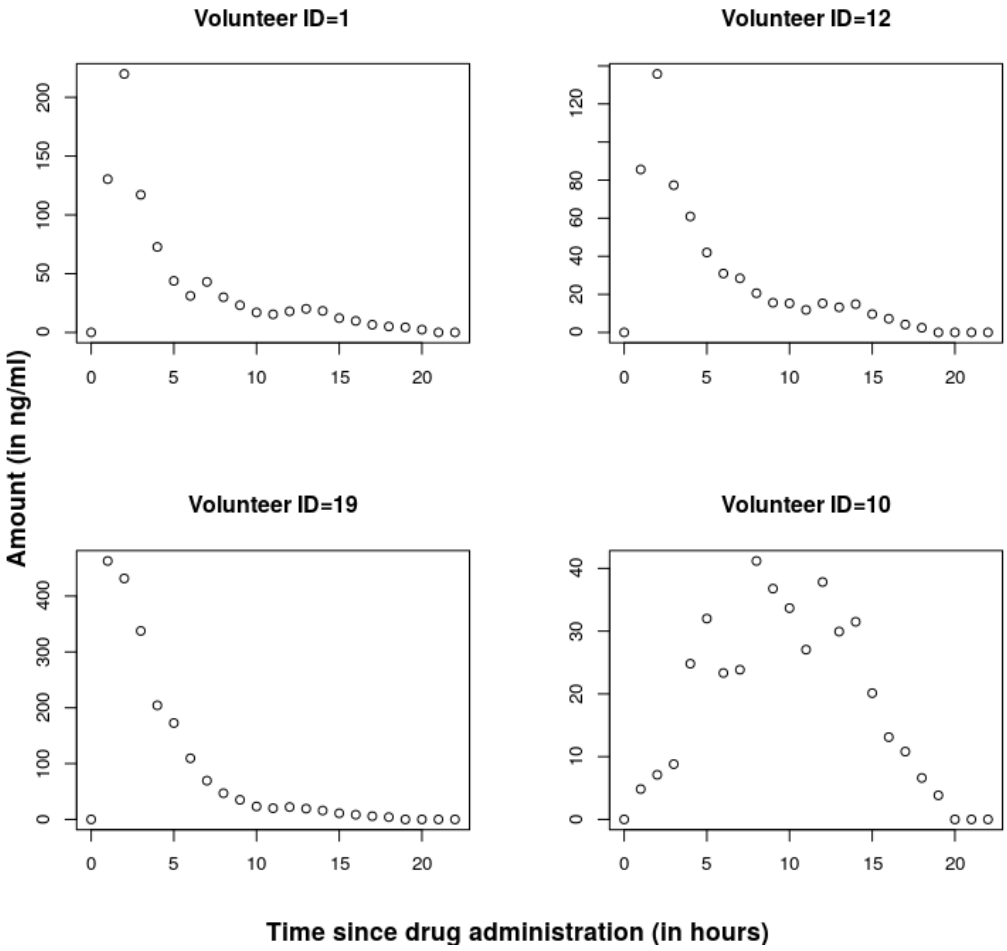
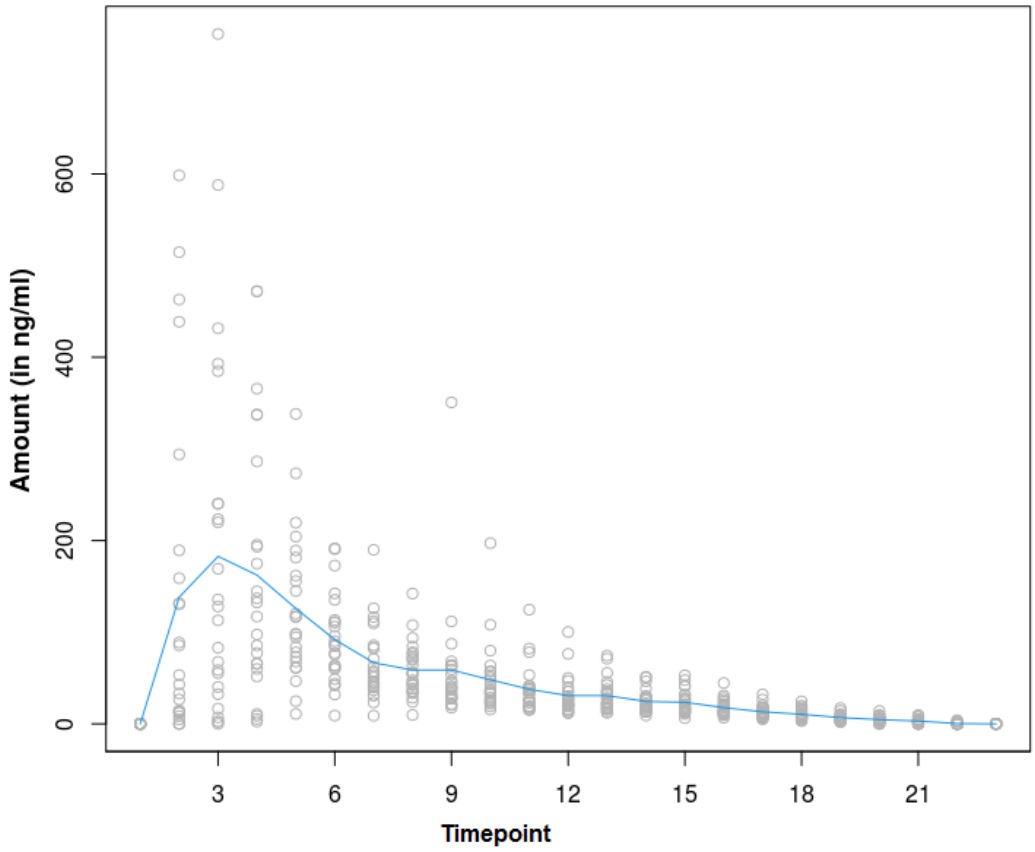


Figure 4.4: **Test drug concentration level for all volunteers**



4.3 Nonlinear models

4.3.1 Model statement

This section presents the joint model formulated to investigate the changing pattern in the drug concentration level in human body in response to the reference and test drugs under consideration.

Let Y_{imt} stands for the i-th (i=1,2,...,24) volunteer's measurement on m-th (m=1,2) longitudinal response at t-th timepoint (sub-cluster) (t=1,2,...,T_i). The vector of responses, Y_{imt} can be represented with, $Y = (Y_{111}, Y_{112}, \ldots, Y_{11T_{23}}, \ldots, Y_{2421}, Y_{2422}, \ldots, Y_{242T_{23}})'$. Two different random effects, subject-specific, U_i and time-specific, V_{imt} are considered in this model. The assumptions of the proposed model are:

Assumption 1: This assumption deals with subject-specific random effects $U_1, U_2, \ldots, U_{24}$. It is assumed that $U_{i'}$s are positive, independently and identically distributed with mean, $E[U_i] = 1$ and variance, $Var[U_i] = \sigma^2$

Assumption 2: This assumption is about time-specific random effects, V_{111}, V_{112}, $\ldots, V_{11T_{23}}, \ldots, V_{2421}, V_{2422}, \ldots, V_{242T_{23}}$. It is assumed that V_{imt}'s are positive and have conditional moment structures depending on subject-specific random effect U as follows,

$$E[V_{imt}|U] = U$$

$$\text{and } Cov[V_{imt}, V_{i'm't'}|U] = \begin{cases} \tau_m^2 \rho_m(t,t') U_i, \text{if } i = i' \text{and} m = m' \\ 0, \text{Otherwise} \end{cases}$$

where the correlation structure among the repeated measures of m-th response, $\rho_{m(t,t')}$ is assumed to be discrete Autoregressive with order 1 (AR (1)).

Assumption 3: This assumption is about the response variables in the model. The responses are assumed to be independent conditional on the random effects U and V. The conditional distribution of the response is assumed to follow Tweedie distribution. The resulting distribution can be written as, $Y_{imt}|(U,V) \sim Tw_{p_m}(\mu_{imt}V_{imt}, \varepsilon_m^2 V_{imt}^{1-p_m})$ where $Tw_{p_m}(\mu_{imt}V_{imt}, \varepsilon_m^2 V_{imt}^{1-p_m})$ denotes the Tweedie distribution with index parameter p, expected value $\mu_{imt}V_{imt}$ and dispersion parameter $\varepsilon_m^2 V_{imt}^{1-p_m}$. This set of exponential distributions can also be termed as the power variance family with,

$$\text{mean, } E[Y_{imt}|(U,V)] = \mu_{imt}V_{imt}$$

$$\begin{aligned} \text{and variance, } Var[Y_{imt}|(U,V)] &= \text{Dispersion Parameter} * (\text{Expected Value})^{p_m} \\ &= \varepsilon_m^2 V_{imt}^{1-p_m} * (\mu_{imt}V_{imt})^{p_m} \\ &= \varepsilon_m^2 V_{imt} \mu_{imt}^{p_m} \end{aligned}$$

where $\mu_{imt} = f(x_{imt}, \beta)$ is a nonlinear function of covariates x_{imt} and regression parameters β.

4.3.2 Mean function

As mean function for both responses, the fitted model used the logarithmic version of the integrated form of the first order compartment model (Willemsen et al., 2017) that can be expressed as follows,

$$\mu_{imt} = e^{(a_m + b_m - c_m)} \frac{\left(e^{-e^{b_m} t} - e^{-e^{a_m} t}\right)}{e^{a_m} - e^{b_m}} \tag{4.1}$$

where

m=1 (parameters related to reference drug), 2 (parameters related to test drug)

t stands for time (in hours) spent since drug administration

a_m is the logarithm of the substance absorption rate

b_m is the logarithm of the substance elimination rate and

c_m is the logarithm of the plasma clearance.

4.3.3 Modeling setups

The concerned response values in the dataset have a lower detection limit 2 ng/ml. The true values of these non-detectable values are unavailable and the actual correlation structure among the repeated measures is also unknown. Considering these facts the following four cases have been investigated in order to test the robustness of the proposed method:

Case-I: Replace the non-detectable values by 0 (minimum) and assume discrete AR (1) correlation structure for repeated measures.

Case-II: Replace the non-detectable values by 2 (maximum) and assume discrete AR (1) correlation structure for repeated measures.

Case-III: Replace the non-detectable values by 0 (minimum) and assume Exchangeable correlation structure for repeated measures.

Case-IV: Replace the non-detectable values by 2 (maximum) and assume Exchangeable correlation structure for repeated measures.

For cases II and IV before fitting the model, 0.00 response values in the dataset were substituted with the lower detection limit (2 ng/ml). After that this limit was subtracted from each of the concentration observations in the dataset. In all four cases the model assumes both jointly modeled outcome variables come from Compound Poisson distribution. This assumption is valid since the data points have mass at zero resulting from the imputation of lower detection limit of concentration level in blood. The rest of the data points are non-negative. As mentioned in Chapter 3, the Tweedie power parameter for Compound Poisson distribution can lie between 1 and 2. A value of 1.6 was set to fit the proposed model by following one of the automated selection approaches available to find Tweedie power parameter that assumes linear relationship between response and time. For both of the responses same estimate (1.6) of the power parameter was found.

4.3.4 Model outputs

The above discussed model can be formulated as an univariate model by considering only one response variable (that means m=1 or 2). The univariate model results in **Table 4.2** suggest that the absorption rate, the elimination rate and plasma clearance resulting from reference drug administration are statistically significantly different from the standard values (a_1=28, b_1=-2, c_1=-3, a_2=33, b_2=-2, c_2=-3) at 5% level of significance. These standard values are found based on graphical inspection and fixed effect nonlinear regression model results. For Case-I and Case-III these standard values are the same since we are using same

imputations for the non-detectable responses. The only difference between the standard values for (Case-I, Case-III) and (Case-II, Case-IV) is the absorption rate parameter for reference drug. It was chosen to be 22 in (Case-II, Case-IV) due to the imputation of non-detectable values as 2. In the Wald-type hypothesis tests (Ma, 1999), the estimates of the multiplicative parameters are compared with standard values. The estimates of random effect parameters indicate a moderate level of subject-specific (0.2516, 0.2116) and time-specific (0.2699, 0.2990) variation in the concerned outcome variables. Another random effect parameter related to the conditional variance of the response given the random effects is resulted to be small in magnitude (0.0873, 0.0359). The measurements taken at different occasions are evident to be strongly (0.6529, 0.8155) correlated to each other.

Table 4.2: **Univariate models for Case-I**

Drug formulation	**Parameter**	**Estimate**	**SE**	**P value**
Reference	a_1	27.9357	9.90×10^{-15}	0.0000
	b_1	-2.0512	0.0125	0.0000
	c_1	-3.8044	0.0109	0.0000
	σ^2	0.2516		
	τ^2	0.2699		
	ε^2	0.0873		
	ρ	0.6529		
Test	a_2	35.3539	1.45×10^{-13}	0.0000
	b_2	-2.2173	0.0546	0.0000
	c_2	-3.9582	0.0522	0.0000
	σ^2	0.2116		
	τ^2	0.2990		
	ε^2	0.0359		
	ρ	0.8155		

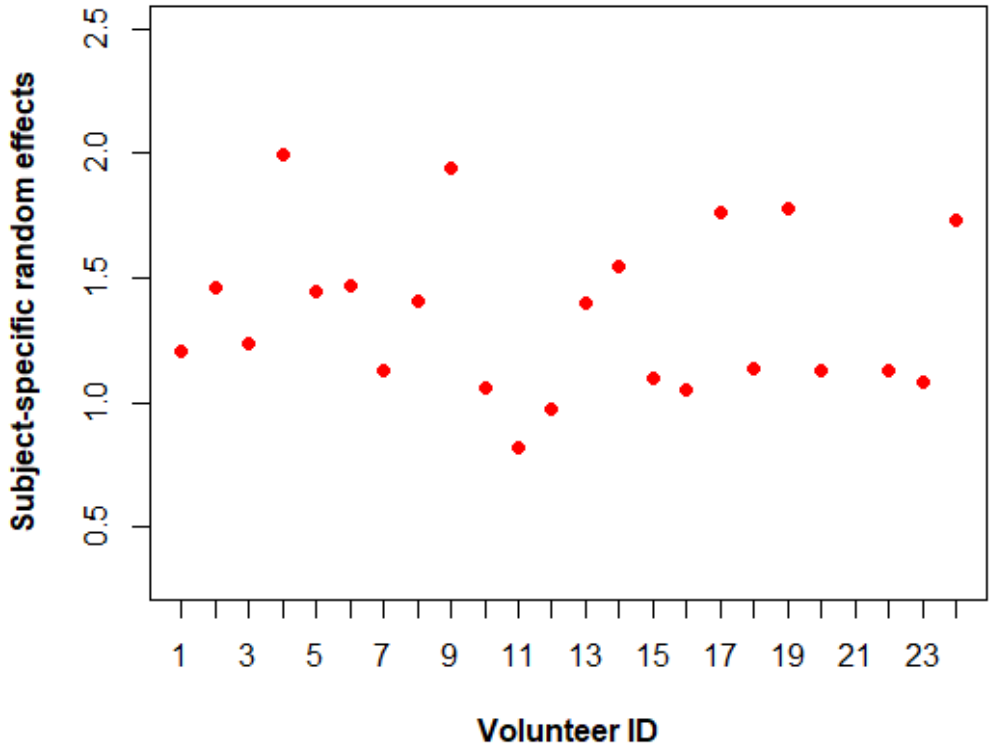
The **Table 4.3** presents the outputs of our proposed joint nonlinear mixed-effects model for Case-I. Overall, the fixed effect parameter estimates from the model are larger than those estimated in the univariate models. This model suggests a relatively smaller subject-specific variation (0.1870) and larger time-specific variations (0.2881, 0.3742) compared to the univariate models. The rest of the random effect parameter estimates are larger than the estimates found in the univariate models.

Table 4.3: **Case-I joint model output**

Parameter	**Estimate**	**SE**	**P value**
a_1	27.9417	4.23×10^{-15}	0.0000
b_1	-2.2615	0.0514	0.0000
c_1	-3.9459	0.0482	0.0000
a_2	33.1900	5.19×10^{-18}	0.0000
b_2	-2.2901	0.0580	0.0000
c_2	-3.9863	0.0542	0.0000
σ^2	0.1870		
${\tau_1}^2$	0.2881		
${\varepsilon_1}^2$	0.0774		
ρ_1	0.7575		
${\tau_2}^2$	0.3742		
${\varepsilon_2}^2$	0.0628		
ρ_2	0.8071		

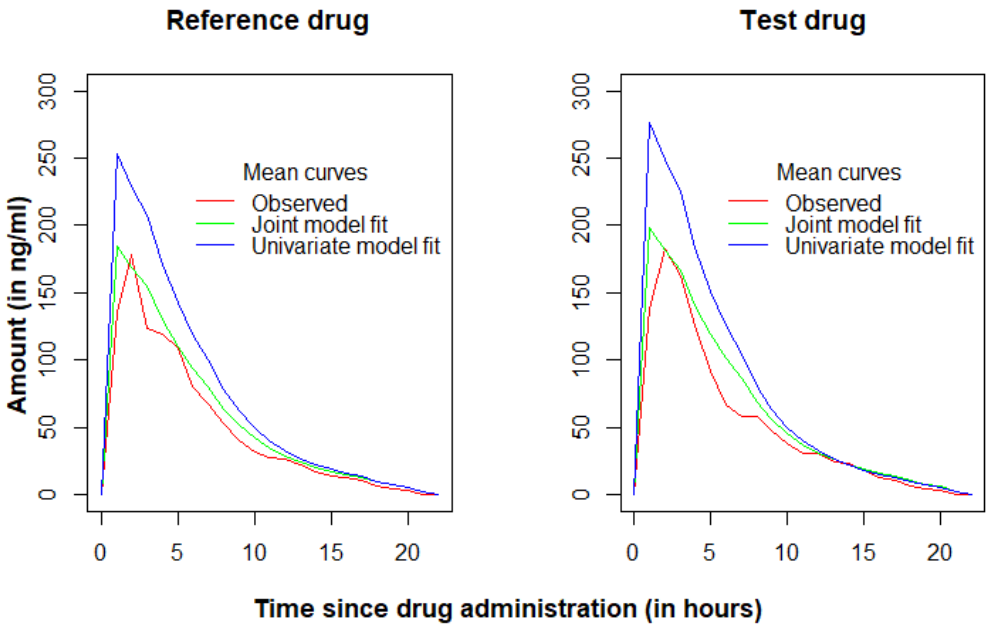
The scatter plot in **Figure 4.5** shows that none of the individuals out of 24 healthy volunteers deviated considerably from the rest of the volunteers. Which supports the presence of a small level of subject-specific variation among individuals.

Figure 4.5: **Subject-specific random effects from Case-I model**



The following graph indicates that the joint model performs better compared to the corresponding univariate models to fit to the observed mean curve for both of the responses under consideration.

Figure 4.6: **Mean curves for Case-I**



The **Table 4.4** and **Table 4.5** present the outputs of univariate and joint non-linear mixed-effects model for Case-II respectively. The fixed effect parameter estimates from the joint model are in line with those from the univariate models. The joint model estimates a relatively smaller subject specific variation and larger time specific variations compared to the univariate models. The rest of the random effect parameter estimates are similar to the estimates found in univariate models.

Table 4.4: **Univariate models for Case-II**

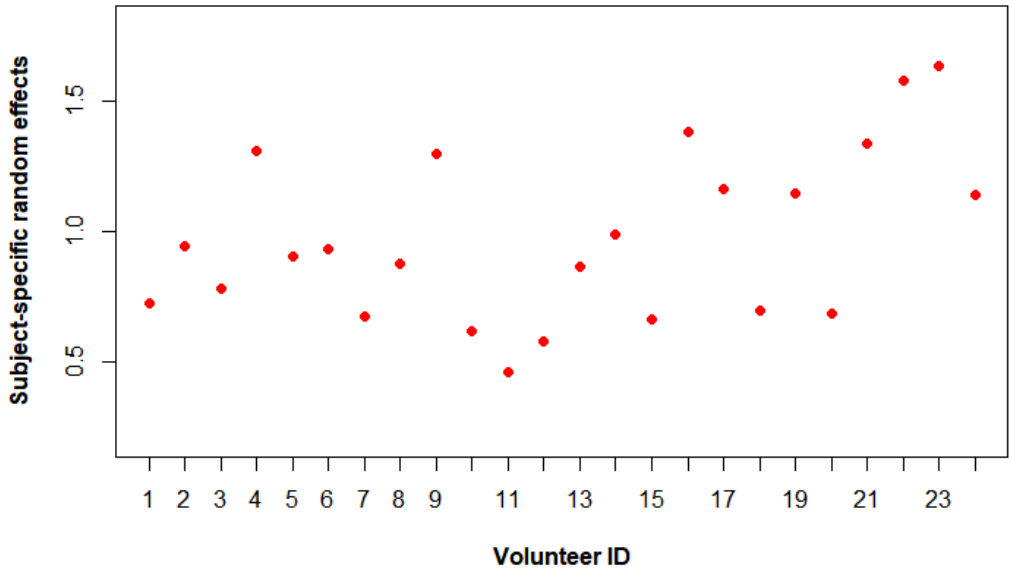
Drug formulation	**Parameter**	**Estimate**	**SE**	**P value**
Reference	a_1	21.9292	1.69×10^{-12}	0.0000
	b_1	-2.2192	0.0487	0.0000
	c_1	-3.9666	0.0436	0.0000
	σ^2	0.3101		
	τ^2	0.1979		
	ε^2	0.0335		
	ρ	0.7862		
Test	a_2	33.1890	1.61×10^{-11}	0.0000
	b_2	-2.1803	0.0517	0.0000
	c_2	-3.9452	0.0495	0.0000
	σ^2	0.2905		
	τ^2	0.2911		
	ε^2	0.0188		
	ρ	0.8326		

Table 4.5: **Case-II joint model output**

Parameter	**Estimate**	**SE**	**P value**
a_1	21.9792	1.66×10^{-12}	0.0000
b_1	-2.2248	0.0450	0.0000
c_1	-3.9573	0.0429	0.0000
a_2	33.1905	4.60×10^{-17}	0.0000
b_2	-2.2493	0.0685	0.0000
c_2	-3.9149	0.0636	0.0079
σ^2	0.2719		
${\tau_1}^2$	0.2187		
${\varepsilon_1}^2$	0.0366		
ρ_1	0.7747		
${\tau_2}^2$	0.4815		
${\varepsilon_2}^2$	0.0289		
ρ_2	0.8203		

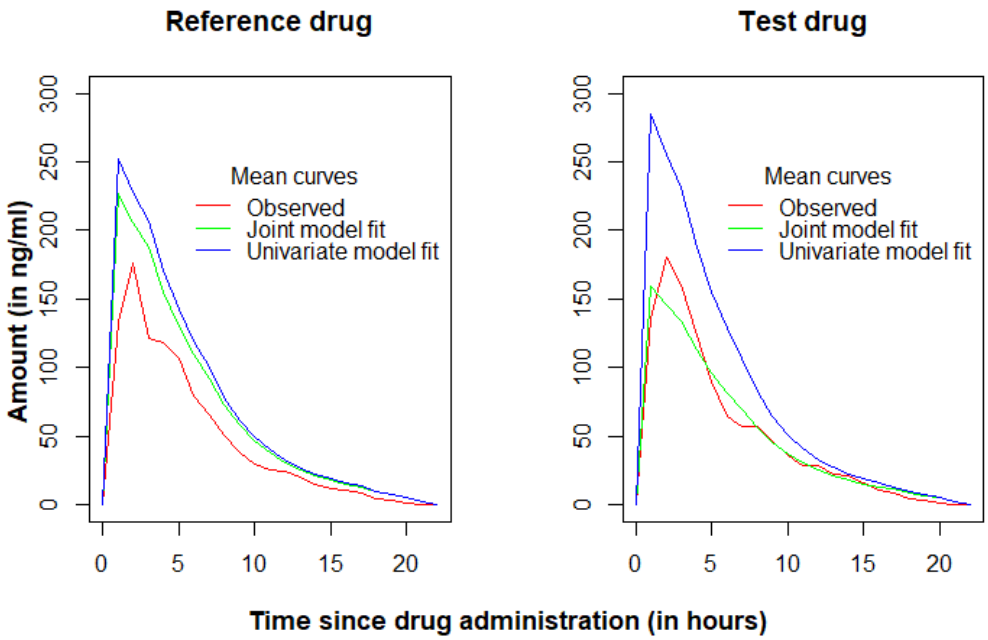
The scatter plot in **Figure 4.7** depicts the presence of a moderate level of subject-specific variation among individuals since none of the individuals significantly deviated from the rest of the individuals.

Figure 4.7: **Subject-specific random effects from Case-II model**



It is evident from the following figure that the difference between univariate model fitted curve and joint model fitted curve is not so high in case of the reference drug response. But the joint model clearly outperforms its univariate counterpart to fit to the observed mean curve for the test drug.

Figure 4.8: **Mean curves for Case-II**



The results from univariate and joint nonlinear mixed-effects models for Case-III are summarized in **Table 4.6** and **Table 4.7** respectively. The joint model produced marginally larger estimates for fixed effects compared to those estimated in univariate models. The joint model estimated a relatively larger time specific variations compared to the univariate models. The rest of the random effect parameter estimates are in line with the estimates from univariate models.

Table 4.6: **Univariate models for Case-III**

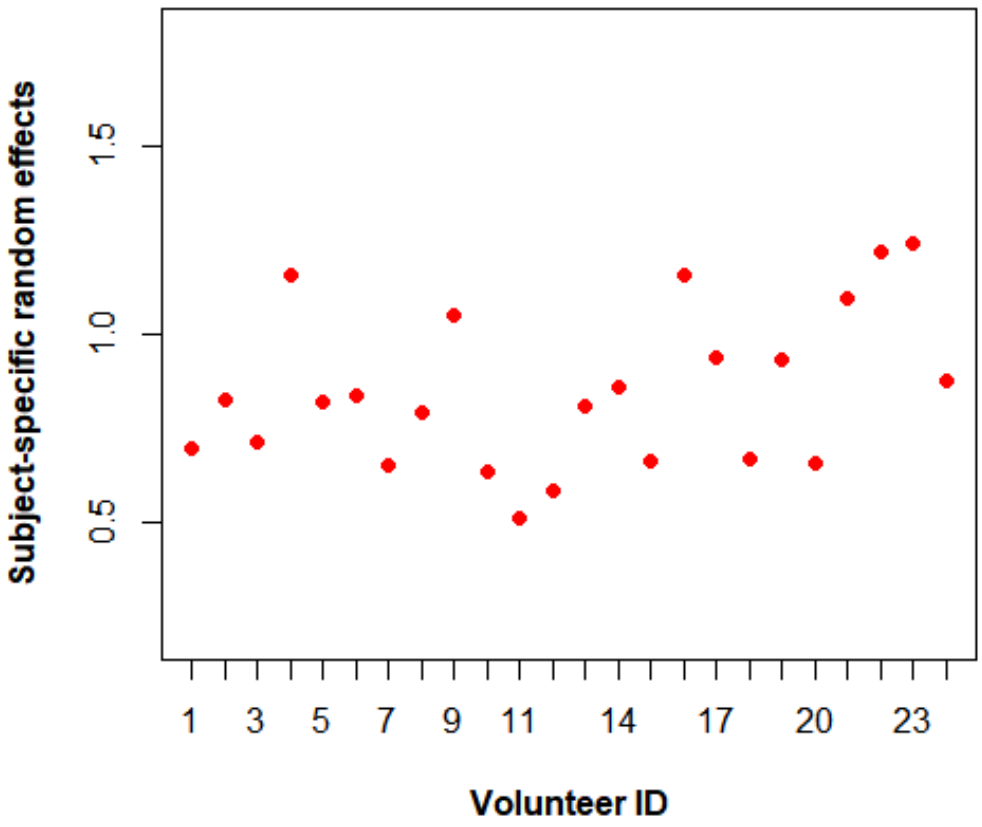
Drug formulation	**Parameter**	**Estimate**	**SE**	**P value**
Reference	a_1	27.9410	3.04×10^{-15}	0.0000
	b_1	-2.2654	0.0292	0.0000
	c_1	-3.9619	0.0337	0.0000
	σ^2	0.0899		
	τ^2	0.2625		
	ε^2	0.1088		
	ρ	0.2551		
Test	a_2	33.2011	3.84×10^{-17}	0.0000
	b_2	-2.2835	0.0320	0.0000
	c_2	-3.9882	0.0366	0.0000
	σ^2	0.0969		
	τ^2	0.3326		
	ε^2	0.0999		
	ρ	0.2257		

Table 4.7: **Case-III joint model output**

Parameter	**Estimate**	**SE**	**P value**
a_1	27.9417	3.11×10^{-15}	0.0000
b_1	-2.3023	0.0297	0.0000
c_1	-3.9942	0.0340	0.0000
a_2	33.1900	2.74×10^{-19}	0.0000
b_2	-2.2897	0.0395	0.0000
c_2	-3.9804	0.0416	0.0000
σ^2	0.0963		
${\tau_1}^2$	0.2832		
${\varepsilon_1}^2$	0.1038		
ρ_1	0.2538		
${\tau_2}^2$	0.5796		
${\varepsilon_2}^2$	0.0927		
ρ_2	0.1925		

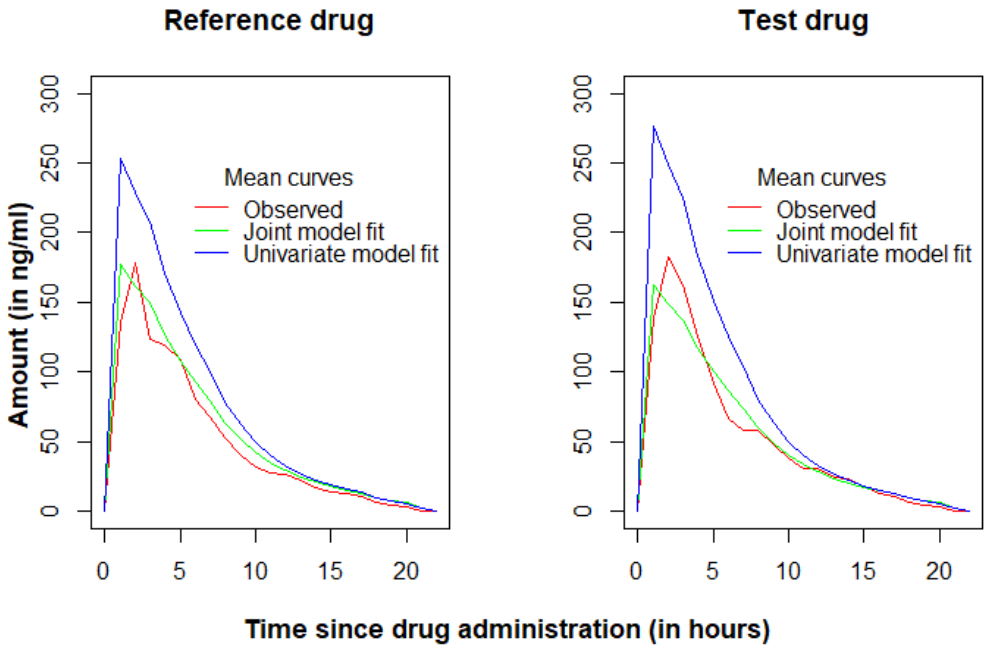
The following scatter plot shows the evidence of a small level of subject-specific variation among individuals since most of estimates of subject-specific random effects lie within a small range (0.5-1.25).

Figure 4.9: **Subject-specific random effects from Case-III model**



It can be seen in the **Figure 4.10** that the joint model clearly outperforms the univariate models to fit to the observed mean curves for both reference drug and test drug responses.

Figure 4.10: **Mean curves for Case-III**



The results for univariate and joint nonlinear mixed-effects models presented in the following tables indicate considerable differences in the estimates for random effect parameters. To illustrate, joint model estimated subject-specific and time-specific variations are smaller than those in the univariate models. Another striking point to notice in the following tables is the large differences in the estimates of the random effect parameter related to the conditional variance for both of the responses. The magnitude of this estimate in the univariate models are 0.0369 and 0.0199 for reference drug and test drug respectively. The same parameter is estimated to be 0.6899 (reference drug) and 1.3797 (test drug) in the joint model. This result is in contrast with the results from the rest of the three cases. On the other hand, fixed effect parameter estimates from the univariate and the joint models are overall similar to each other.

Table 4.8: **Univariate models for Case-IV**

Drug formulation	**Parameter**	**Estimate**	**SE**	**P value**
Reference	a_1	21.9791	1.66×10^{-12}	0.0000
	b_1	-2.2413	0.0441	0.0000
	c_1	-3.9942	0.0429	0.0000
	σ^2	0.3260		
	τ^2	0.1871		
	ε^2	0.0361		
	ρ	0.7903		
Test	a_2	33.1905	5.46×10^{-17}	0.0000
	b_2	-2.2043	0.0505	0.0000
	c_2	-3.9749	0.0485	0.0000
	σ^2	0.3083		
	τ^2	0.2740		
	ε^2	0.0199		
	ρ	0.8327		

Table 4.9: **Case-IV joint model output**

Parameter	**Estimate**	**SE**	**P value**
a_1	21.9792	1.46×10^{-12}	0.0000
b_1	-2.2512	0.0299	0.0000
c_1	-3.9885	0.0378	0.0000
a_2	33.1905	1.38×10^{-17}	0.0000
b_2	-2.2652	0.0339	0.0000
c_2	-3.9692	0.0393	0.0005
σ^2	0.1700		
${\tau_1}^2$	0.0816		
${\varepsilon_1}^2$	0.6899		
ρ_1	0.6758		
${\tau_2}^2$	0.0253		
${\varepsilon_2}^2$	1.3797		
ρ_2	0.8700		

The **Figure 4.11** portrays the level of variation among volunteers. The **Figure 4.12** presents a comparison between univariate and joint model fits to the observed mean curves for the responses. Both of these figures have similar pattern as it was found in the Case-II model results. Small level of subject-specific variation is evident in **Figure 4.11**. Overall, the joint model results for Case-IV fits the observed mean curve better than the univariate models.

Figure 4.11: **Subject-specific random effects from Case-IV model**

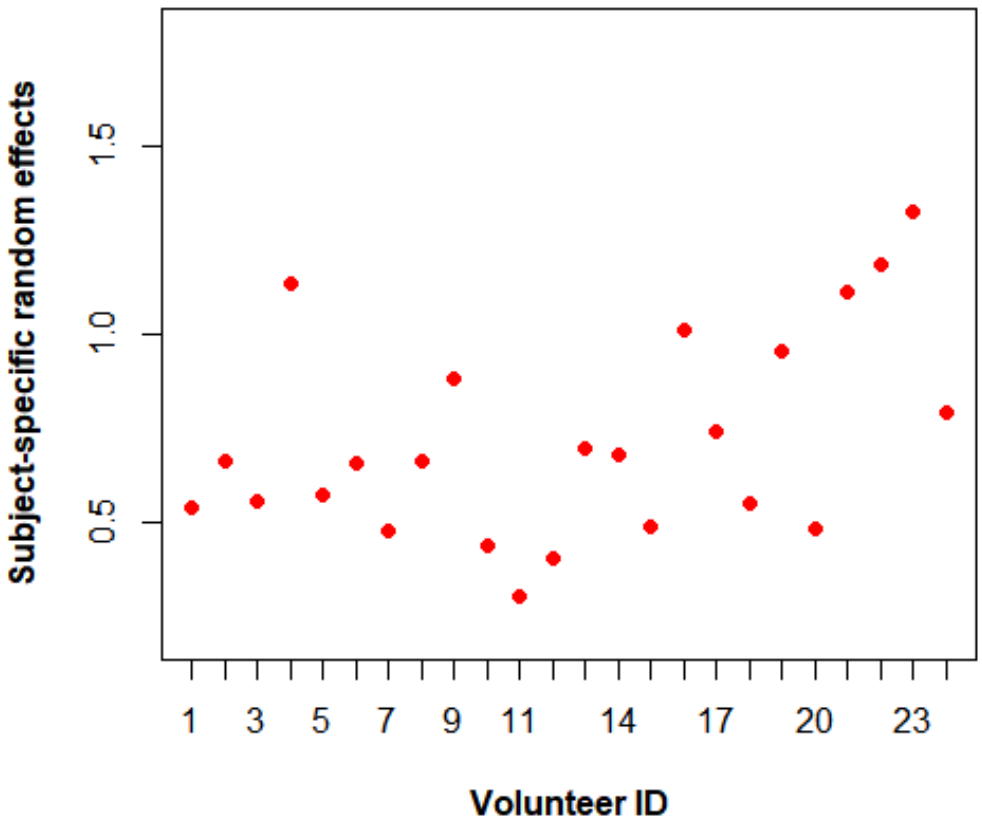
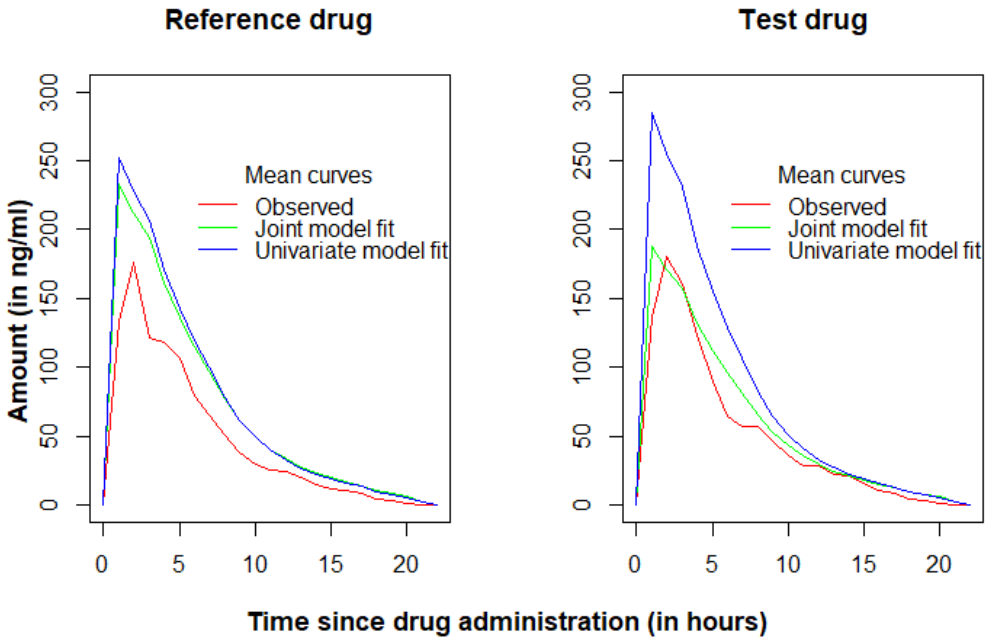


Figure 4.12: **Mean curves for Case-IV**



In order to assess the bioequivalence between the two formulations under study, simple Z-tests of difference between the estimated fixed effect parameters are conducted from the fitted joint model outputs. The summarized results of these tests are provided in **Table 4.10**:

Table 4.10: **Test of difference in fixed effect parameter estimates**

Case	Difference	SE	Z	P-value
I	$\hat{a_1}$-$\hat{a_2}$=-5.2483	1.65×10^{-12}	<-318.995	0.0000
	$\hat{b_1}$-$\hat{b_2}$=0.0286	0.0689	0.4152	0.6780
	$\hat{c_1}$-$\hat{c_2}$=0.0404	0.0619	0.6524	0.5141
II	$\hat{a_1}$-$\hat{a_2}$=-11.2113	4.23×10^{-15}	<-264.471	0.0000
	$\hat{b_1}$-$\hat{b_2}$=0.0245	0.0698	0.3509	0.7256
	$\hat{c_1}$-$\hat{c_2}$=-0.0424	0.0615	-0.6892	0.4906
III	$\hat{a_1}$-$\hat{a_2}$=-5.2483	1.46×10^{-12}	<-169.0821	0.0000
	$\hat{b_1}$-$\hat{b_2}$=-0.0126	0.0424	-0.2971	0.7663
	$\hat{c_1}$-$\hat{c_2}$=-0.0138	0.0437	-0.0751	0.7522
IV	$\hat{a_1}$-$\hat{a_2}$=-11.2113	3.10×10^{-15}	<-766.9054	0.0000
	$\hat{b_1}$-$\hat{b_2}$=0.0140	0.0367	0.3807	0.7034
	$\hat{c_1}$-$\hat{c_2}$=-0.0123	0.0368	-0.3345	0.7379

The results in the above table suggest that the substance absorption rate parameter for reference drug and test drug are significantly different in all four investigated cases. The differences between the important parameters related to substance elimination rate and plasma clearance are revealed not to be significant in all the cases. The substance absorption rate determines the steepness of the initial slope of the mean curve. The larger substance elimination rate results in sharp decrease in the later stage and reduced initial slope of the curve. The large value of plasma clearance parameter stretches the mean curve upward (Willemsen et al., 2017). Considering the influence of the parameters on the mean curve, we can conclude that the above test results show evidence of potential bioequivalence between the reference drug and the test drug formulations. The similarity in the test results from all four cases indicates the consistency of the proposed data analysis method.

Depending on the results from the above table, the joint models in Case-I, II, III and IV can be simplified. If we set b_1=b_2 and c_1=c_2, then a simplified

version of the mean function can be written as follows,

$$\mu_{imt} = e^{(a_m+b-c)} \frac{\left(e^{-e^b t} - e^{-e^{a_m} t}\right)}{e^{a_m} - e^b} \tag{4.2}$$

where the variable t and the parameters bear the same meaning as the mean function introduced in 4.1.

Chapter 5

Simulation Study: An Evaluation of the Proposed Modeling Framework

This chapter introduces simulation studies that involved generating data by random sampling from known Tweedie distributions. We conducted simulation experiments to check if our proposed nonlinear modeling approach works in the scenarios for which it is intended to deal with. The true values for the assumed model parameters are calculated by fitting our proposed joint model to Losartan data in four different cases introduced in the previous chapter. Pseudo-random data points were generated based on the resulting estimates of parameters from the fitted joint models. Two performance measures: bias and standard error (SE) for the model parameters have been reported from all the simulation experiments. Simulated and Estimated standard errors are calculated for each of the simulation experiment. Simulated standard errors are the standard deviation of the estimates and Estimated standard errors are the mean of the standard errors of the fixed effects from the joint models in an experiment.

The rest of the parts of this chapter present the simulation experiments in detail with different components including data generating mechanism and per-

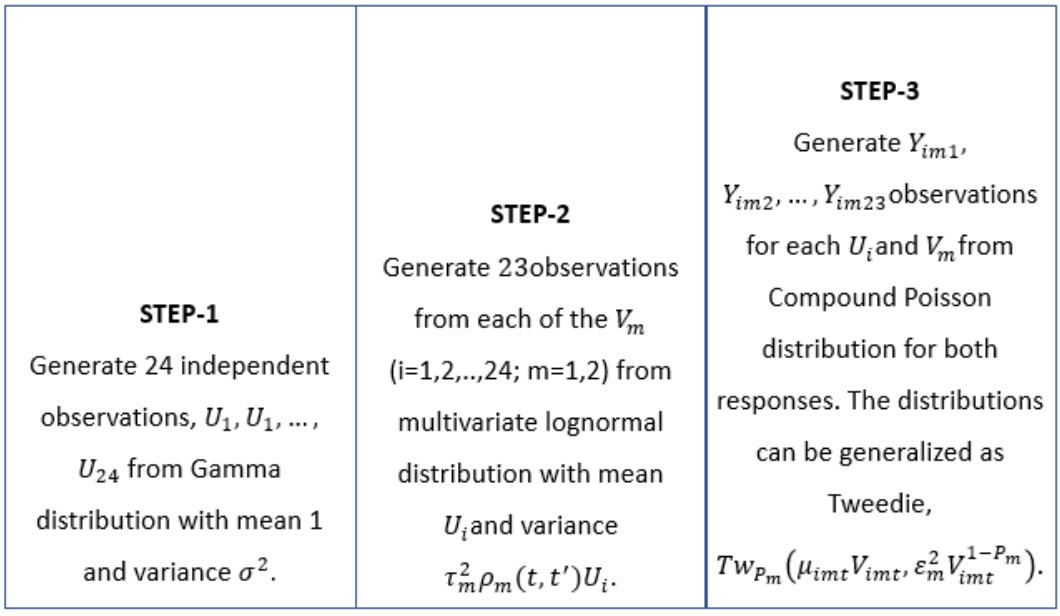
formance measures.

The dataset and the simplified mean function assumed to fit the noinlinear models for Losartan data are introduced in the previous chapter. For this dataset we conducted four simulation experiments based on the Case I, Case II, Case III and Case IV joint models presented in Chapter 4.

5.1 Data generating mechanism

The proposed fitting algorithm is run for 500 randomly generated sample of responses. The random data generation mechanism for joint model experiments are exhibited in **Figure 5.1**.

Figure 5.1: **Random data generation from joint model**



5.2 Simulation results for Case-I model

On average it took 31 iterations to converge for each of the simulated samples in the Case-I model simulation experiment. The algorithm converged for a total of 407 (81.4% of 500) out of 500 runs. The estimated values are averages from the converged samples. A summary of the results of this experiment is

organized in **Table 5.1**:

Table 5.1: **Case-I model based simulation results**

Parameter	**True value**	**Estimate**	**Bias**	**SE**	
				Simulated	**Estimated**
a_1	27.9417	27.0944	-0.8473	0.8637	6.74×10^{-06}
b	-2.2615	-2.1653	0.0962	0.0298	0.0025
c	-3.9459	-3.7927	0.1532	0.4877	0.0022
a_2	33.1900	33.3165	0.1265	0.7023	1.21×10^{-27}
σ^2	0.1870	0.2159	0.0289	0.0234	
${\tau_1}^2$	0.2881	0.2601	-0.0280	0.0208	
${\varepsilon_1}^2$	0.0774	0.0563	-0.0211	0.0106	
ρ_1	0.7575	0.7655	0.0080	0.0075	
${\tau_2}^2$	0.3742	0.3652	-0.0090	0.0145	
${\varepsilon_2}^2$	0.0628	0.0564	-0.0064	0.0078	
ρ_2	0.8071	0.8159	0.0088	0.0097	

5.3 Simulation results for Case-II model

In this experiment on the fitted joint model 346 (69.2% of 500) sample converged for the joint nonlinear algorithm. On average it took around 42 iterations for the algorithm to converge in this simulation experiment. A summary of this experiment is presented in **Table 5.2**:

Table 5.2: **Case-II model based simulation results**

Parameter	**True value**	**Estimate**	**Bias**	**SE**	
				Simulated	**Estimated**
a_1	21.9792	21.0232	-0.9560	0.4137	1.07×10^{-24}
b	-2.2248	-2.1766	0.0482	0.0112	0.0044
c	-3.9573	-3.9562	0.0011	0.0932	0.0038
a_2	33.1905	32.8583	-0.3322	1.1583	2.65×10^{-11}
σ^2	0.2719	0.2358	-0.0361	0.0240	
${\tau_1}^2$	0.2187	0.2560	0.0373	0.0176	
${\varepsilon_1}^2$	0.0366	0.0541	0.0175	0.0100	
ρ_1	0.7747	0.7669	-0.0078	0.0085	
${\tau_2}^2$	0.4815	0.5219	0.0404	0.0811	
${\varepsilon_2}^2$	0.0289	0.0399	0.0110	0.0081	
ρ_2	0.8203	0.8164	-0.0039	0.0089	

5.4 Simulation results for Case-III model

The joint nonlinear algorithm converged for 356 (71.2% of 500) samples in this simulation experiment. The average number of iterations for the algorithm to converge in this simulation experiment is around 18. The summarized results of this exiperiment is presented in **Table 5.3**:

Table 5.3: **Case-III model based simulation results**

Parameter	**True value**	**Estimate**	**Bias**	**SE**	
				Simulated	**Estimated**
a_1	27.9417	26.4881	-1.4563	0.7112	2.10×10^{-07}
b	-2.3023	-2.2003	0.1020	0.0768	0.0072
c	-3.9942	-3.9903	0.0039	0.0626	0.0064
a_2	33.1900	33.7063	0.5163	0.4171	3.24×10^{-18}
σ^2	0.0963	0.1182	0.0219	0.0308	
${\tau_1}^2$	0.2832	0.2741	-0.0091	0.0199	
${\varepsilon_1}^2$	0.1038	0.0879	-0.0159	0.0068	
ρ_1	0.2538	0.2609	0.0071	0.0027	
${\tau_2}^2$	0.5796	0.5510	-0.0286	0.0398	
${\varepsilon_2}^2$	0.0927	0.0783	-0.0144	0.0072	
ρ_2	0.1425	0.1504	0.0079	0.0010	

5.5 Simulation results for Case-IV model

In this experiment on the fitted joint model 369 (73.8% of 500) sample converged for the joint nonlinear algorithm. The algorithm converged with an average of 48 iterations in this simulation experiment. A summary of this exiperiment is presented in **Table 5.4**:

Table 5.4: **Case-IV model based simulation results**

Parameter	**True value**	**Estimate**	**Bias**	**SE**	
				Simulated	**Estimated**
a_1	21.9792	20.5170	-0.9560	0.5149	3.93×10^{-13}
b	-2.2512	-2.2398	0.0482	0.0181	0.0080
c	-3.9885	-3.9499	0.0011	0.0745	0.0023
a_2	33.1905	32.9379	-0.3322	1.0444	1.06×10^{-17}
σ^2	0.1700	0.1487	-0.0361	0.0253	
${\tau_1}^2$	0.0861	0.0946	0.0373	0.0172	
${\varepsilon_1}^2$	0.6899	0.6974	0.0175	0.0847	
ρ_1	0.6758	0.6619	-0.0078	0.0035	
${\tau_2}^2$	0.0253	0.0346	0.0404	0.0287	
${\varepsilon_2}^2$	1.3757	1.3886	0.0110	0.0156	
ρ_2	0.8700	0.8574	-0.0039	0.0021	

The results in the above tables suggest that in general the fixed effect parameter estimates have large bias and SE in all the four experiments. It is not a rare situation to encounter in nonlinear models (Cook et al., 1986). On the other hand, the random effect parameter estimates have shown relatively small bias and SE compared to the fixed effect parameter estimates. This result is in line with an earlier study that fitted nonlinear univariate model of this type (Snow, 2019). The simulated SE for the fixed effect parameters are considerably larger compared to Estimated SE, particularly for substance absorption rate parameter in all the cases. One of the potential reasons behind this result could be the sensitivity of the nonlinear regression models to the fixed effect parameters in genereal (Snow, 2019).

Chapter 6

Conclusion

We proposed a joint Generalized Nonlinear Mixed-Effects Modeling framework designed for Longitudinal Data. The Joint model is developed by following a shared random effects approach.

The Newton–Raphson algorithm is implemented to find the estimates for the parameters in the model. The bouncing algorithm procedure is employed to deal with the absurd estimate and non-convergence issues that usually arise in nonlinear models. Subject-specific random effects are included in the model to capture the unobserved heterogeneity among subjects of interest. Time-specific random effects conditional on the subject-specific random effects are introduced to account for the correlation among the repeated measures for each response. The random effects are estimated with orthodox best linear unbiased predictors (BLUPs).

The proposed model can accommodate response variables belonging to Tweedie exponential distributions. It also allows for mixed (such as discrete and continuous) types of responses to be modeled together. The distributional assumption about the random effects is not restricted to normal. It requires only the first two moments of the random effects to be specified to optimize the parameter estimation process. A wide range of correlation structures is acceptable in the model to explore the dependencies among repeated measures within subjects.

There is still room for improvement in the proposed modeling framework. Fu-

ture work may include the development of a robust prediction interval for the outcome variables in the model. As a medium of model fitting performance, one can look for approaches beyond simulation and simple graphical assessment. A considerable amount of missing values is a common problem in longitudinal studies. The quality of the framework can be upgraded by addressing this particular issue in this line of research. It is possible to estimate standard error for dispersion and correlation parameters by implementing resampling algorithms such as Bootstrapping and Jackknife resampling.

Bibliography

Aschengrau, A. and Seage, G. R. (2013). *Essentials of epidemiology in public health*. Jones & Bartlett Publishers.

Clark, D. R. and Thayer, C. A. (2004). A primer on the exponential family of distributions. In *Casualty Actuarial Society Spring Forum*, pages 117–148.

Cook, R. D., Tsai, C.-L., and Wei, B. (1986). Bias in nonlinear regression. *Biometrika*, 73(3):615–623.

Davidian, M. and Giltinan, D. M. (2003). Nonlinear models for repeated measurement data: an overview and update. *Journal of agricultural, biological, and environmental statistics*, 8(4):387.

Deisboeck, T. and Kresh, J. Y. (2007). *Complex systems science in biomedicine*. Springer Science & Business Media.

Diggle, P., Diggle, P. J., Heagerty, P., Liang, K.-Y., Heagerty, P. J., Zeger, S., et al. (2002). *Analysis of longitudinal data*. Oxford University Press.

Dronda, F., Moreno, S., Moreno, A., José, L. C., Pérez-Elías, M. J., and Antela, A. (2002). Long-term outcomes among antiretroviral-naive human immunodeficiency virus–infected patients with small increases in cd4+ cell counts after successful virologic suppression. *Clinical infectious diseases*, 35(8):1005–1009.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2012). *Applied longitudinal analysis*, volume 998. John Wiley & Sons.

Fitzmaurice, G. M. and Ravichandran, C. (2008). A primer in longitudinal data analysis. *Circulation*, 118(19):2005–2010.

Grimm, K. J. (2019). Review of intensive longitudinal methods: An introduction to diary and experience sampling research.

Hu, C. and Sale, M. E. (2003). A joint model for nonlinear longitudinal data with informative dropout. *Journal of Pharmacokinetics and Pharmacodynamics*, 30(1):83–103.

Huang, Y. and Dagne, G. (2012). Bayesian semiparametric nonlinear mixed-effects joint models for data with skewness, missing responses, and measurement errors in covariates. *Biometrics*, 68(3):943–953.

Jørgensen, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 49(2):127–145.

Jorgensen, B. (1997). *The theory of dispersion models*. CRC Press.

Lane, P. W. (1996). Generalized nonlinear models. In *COMPSTAT*, pages 331–336. Springer.

Lee, Y., Nelder, J. A., et al. (2004). Conditional and marginal models: another view. *Statistical Science*, 19(2):219–238.

Li, J. (2017). *Joint Tweedie Mixed Models for Longitudinal Data of Mixed Types*. PhD thesis, University of New Brunswick.

Linnhoff, S., Smith, K. T., and Smith, L. M. (2020). An examination of longitudinal study typologies for business research. *International Journal of Business and Globalisation*, 24(2):212–239.

Lu, X. and Huang, Y. (2014). Bayesian analysis of nonlinear mixed-effects mixture models for longitudinal data with heterogeneity and skewness. *Statistics in medicine*, 33(16):2830–2849.

Lu, X., Huang, Y., and Zhu, Y. (2016). Finite mixture of nonlinear mixed-effects joint models in the presence of missing and mismeasured covariate, with application to AIDS studies. *Computational Statistics & Data Analysis*, 93:119–130.

Lucia, S., Andersson, J. A., Brandt, H., Diehl, M., and Engell, S. (2014). Handling uncertainty in economic nonlinear model predictive control: A comparative case study. *Journal of Process Control*, 24(8):1247–1259.

Ma, R. (1999). *An orthodox BLUP approach to generalized linear mixed models*. PhD thesis, University of British Columbia.

Ma, R. and Jørgensen, B. (2007). Nested generalized linear mixed models: an orthodox best linear unbiased predictor approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):625–641.

Ma, R., Yan, G., and Hasan, M. T. (2018). Tweedie family of generalized linear models with distribution-free random effects for skewed longitudinal data. *Statistics in medicine*, 37(24):3519–3532.

Mathalon, D. H., Ford, J. M., and Pfefferbaum, A. (2000). Trait and state aspects of p300 amplitude reduction in schizophrenia: a retrospective longitudinal study. *Biological psychiatry*, 47(5):434–449.

Muff, S., Held, L., and Keller, L. F. (2016). Marginal or conditional regression models for correlated non-normal data? *Methods in Ecology and Evolution*, 7(12):1514–1524.

Overall, J. E. and Tonidandel, S. (2004). Robustness of generalized estimating equation (gee) tests of significance against misspecification of the error structure model. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 46(2):203–213.

Pillai, G. C., Mentré, F., and Steimer, J.-L. (2005). Non-linear mixed effects modeling–from methodology and software development to driving implemen-

tation in drug development science. *Journal of pharmacokinetics and pharmacodynamics*, 32(2):161–183.

Proust-Lima, C., Letenneur, L., and Jacqmin-Gadda, H. (2007). A nonlinear latent class model for joint analysis of multivariate longitudinal data and a binary outcome. *Statistics in medicine*, 26(10):2229–2245.

Rosen, M. J., Bauer, J. J., Harmaty, M., Carbonell, A. M., Cobb, W. S., Matthews, B., Goldblatt, M. I., Selzer, D. J., Poulose, B. K., Hansson, B. M., et al. (2017). Multicenter, prospective, longitudinal study of the recurrence, surgical site infection, and quality of life after contaminated ventral hernia repair using biosynthetic absorbable mesh: the COBRA study. *Annals of surgery*, 265(1):205.

Ross, S. M. et al. (2006). *A first course in probability*, volume 7. Pearson Prentice Hall Upper Saddle River, NJ.

Sabzpoushan, S. (2020). A flexible nonlinear model for simulating growth systems. *Communications in Nonlinear Science and Numerical Simulation*, 82:105009.

Singer, J. D., Willett, J. B., Willett, J. B., et al. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford university press.

Snow, A. (2019). *Tweedie Nonlinear Mixed Models for Longitudinal Data*. PhD thesis, UNIVERSITY OF NEW BRUNSWICK.

Tang, N.-S. and Zhao, H. (2014). Bayesian analysis of nonlinear reproductive dispersion mixed models for longitudinal data with nonignorable missing covariates. *Communications in Statistics-Simulation and Computation*, 43(6):1265–1287.

Taris, T. W. (2000). *A primer in longitudinal data analysis*. Sage.

Tweedie, M. C. (1984). An index which distinguishes between some important exponential families. In *Statistics: Applications and new directions: Proc. Indian statistical institute golden Jubilee International conference*, volume 579, pages 579–604.

Verbeke, G. and Davidian, M. (2009). Joint models for longitudinal data: Introduction and overview.

Vonesh, E. F. (1992). Non-linear models for the analysis of longitudinal data. *Statistics in medicine*, 11(14-15):1929–1954.

Weiss, N. A. (2005a). A course in probability. 2005.

Weiss, R. E. (2005b). *Modeling longitudinal data*. Springer Science & Business Media.

Willemsen, S. P., Russo, C. M., Lesaffre, E., and Leão, D. (2017). Flexible multivariate nonlinear models for bioequivalence problems. *Statistical Modelling*, 17(6):449–467.

Wolff, J. J., Swanson, M. R., Elison, J. T., Gerig, G., Pruett, J. R., Styner, M. A., Vachet, C., Botteron, K. N., Dager, S. R., Estes, A. M., et al. (2017). Neural circuitry at age 6 months associated with later repetitive behavior and sensory responsiveness in autism. *Molecular autism*, 8(1):8.

Wu, H. and Ding, A. A. (1999). Population hiv-1 dynamics in vivo: applicable models and inferential tools for virological data from aids clinical trials. *Biometrics*, 55(2):410–418.

Wu, L. (2004). Exact and approximate inferences for nonlinear mixed-effects models with missing covariates. *Journal of the American Statistical Association*, 99(467):700–709.

Wu, L. (2009). *Mixed effects models for complex data*. CRC Press.

Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, pages 121–130.

Vita

Candidate's full name:
Md Ashiqul Haque

University attended :
Master of Science, 2020, University of New Brunswick, Fredericton, NB, Canada
Master of Science (course based), 2018, Shahjalal University of Science and Technology, Sylhet, Bangladesh
Bachelor of Science, 2017, Shahjalal University of Science and Technology, Sylhet, Bangladesh

Conference Presentations:
Haque, Ashiqul; Hasan, Tariq; Ma, Renjun (2020). Nonlinear Joint Models for Longitudinal Data.
The 2020 online Canadian Statistics Student Conference.