A Joint Mixed Model for Clustered Binary and Continuous Outcomes

by

Tianyi Xia

Bachelor of Science, University of Toronto, 2015

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):	Guohua Yan, Ph.D, Statistics
	Renjun Ma, Ph.D, Statistics
Examining Board:	Tariq Hasan, Ph.D, Statistics
	Fan-Rui Meng, Ph.D, Forestry & Environmental Management

This thesis is accepted by the

Dean of Graduate Studies

THE UNIVERSITY OF NEW BRUNSWICK

Sepmeter, 2020

Abstract

Recently, statisticians are very interested in developing new methods for the data with different responses types. The generalized linear model (GLM) can analyze different responses separately. The generalized linear mixed model (GLMM), an extension to the GLM, can consider both fixed effects and random effects. However, the difficulty lies in how to analyze two different responses at the same time, while considering fixed effects, random effects and the correlation between the responses.

In this thesis, we proposed a model through the GLMM based on Tweedie distribution and combine it with Bernoulli distribution to analyze simultaneously clustered binary responses and continuous response. The proposed model is very flexible as the Tweedie model can handle different distributions with different power index parameters. We used the best linear unbiased predictor (BLUP) in our model to provide optimal parameter estimates results. We analyzed the toxicity study of ethylene glycol (EG) data to demonstrate the performance of our proposed model. We also conducted a simulation study to assess the overall model performance.

Dedication

This thesis is dedicated to my parents, who supported me with endless love and encouragement throughout my education.

Acknowledgements

I would like to express my heartiest gratitude to my supervisor Dr.Guohua Yan and Dr.Renjun Ma for his guidance, patience and kindness. I would like to thank Dr.Guohua Yan for his helpful instruction, excellent guidance and great patience. I would like to thank Dr.Renjun Ma for inspiring me a new way of learning and thinking as well as his valuable guidance. Without their advice, I might not have completed this thesis.

I would like to thank Dr.Jeffery Picka for his helpful instruction and incredible patience during my graduate seminar.

I would also like to thank Dr.Tariq Hasan for his kindness and support.

In addition, I would like to thank the faculty in UNB Mathematics and Statistics department for providing me an excellent academic program and friendly study environment.

Finally, I would like to thank all my friends for their support and encouragement.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Today, there are many ways to analyze data of different types. The most basic one, the simple linear regression can analyze a continuous response with only one explanatory variable. A more complicated one, the generalized linear model (GLM) can study a relationship between a continuous or discrete response and one or more explanatory variables. Unlike linear models, the generalized linear models include a variety of models that includes normal, binomial, Poisson and multinomial as special cases (Jiang, 2007). But one limitation of the generalized linear model is that the linear predictor includes only fixed effects (Hedeker, 2005). The generalized linear mixed model (GLMM) is an extension to the generalized linear model in which the linear predictor is able to include both random effects and fixed effects (Breslow et al.,1993). The random effects allow the correlation

between the clustered measurements to be incorporated into the estimates of parameters, standard errors, interval estimates, and tests of hypotheses (Gibbons et al., 2010). Although there are many statistical methods, researchers often face some difficulties when dealing with the data containing mixed responses. For example, Fitzmaurice et al. (2011) used the multilevel linear model to analyze the continuous response of the Developmental Toxicity Study of Ethylene Glycol dataset and use the multilevel generalized linear model to analyze the binary response of the Developmental Toxicity Study of Ethylene Glycol dataset. Although multilevel linear models can be regarded as extensions of the linear mixed effects models which allow random effects to be incorporated at more than one level, it cannot analyze the two different responses together. In other words, the multilevel linear model can only analyze the two different responses separately. However, when two outcomes are measured in the same unit, they are likely to be correlated. Therefore, we should analyze them together. Ezzalfani et al. (2018) have developed joint modelling of a binary and a continuous outcome. However, their research was based on non-clustered data which means that there is no random effect between different clusters. In addition, some researchers have developed the regression models for a bivariate discrete and continuous outcome with clustering, more details can be seen from Fitzmaurice et al. (1995) and Lin et al. (2010).

In this thesis, the model will build on the concept of previous work done by Jørgensen. (1987), Ma. (1999) and Ma et al. (2007). The structure of the model is based on Tweedie generalized linear mixed model where the

responses are assumed to follow the Tweedie exponential dispersion distributions. The model is able to handle clustered responses with binary and continuous types and analyze these different responses altogether while incorporating the random effects. Compare with the joint model which have developed by Ezzalfani et al. (2018), our model will include both cluster-level random effects and subject-level random effects. The best linear unbiased predictor (BLUP) is also used in our model to provide optimal parameter estimates results.

1.2 Organization of the thesis

Chapter 2 will mainly focus on an overreview of previous work and preexisting methods. We will briefly introduce the generalized linear model (GLM), Tweedie exponential dispersion model and quasi likelihood function. The advantage and disadvantage as well as the limitation of each method will be also discussed in chapter 2.

Chapter 3 will introduce the proposed model based on three assumptions. The details of the three assumptions can be found in section 3.1. The moment of structure will be also included and can be found in section 3.2. Section 3.3 will introduce the best linear unbiased predictors (BLUP) of random effects which are provide the best optimal parameter estimates results. Estimation of parameters will be developed in section 3.4. The parameters initialization process and updating iteration process will be discussed in the end of chapter 3.

Chapter 4 will use the proposed model to analyze the development toxicity study of ethylene glycol (EG) data. The background information about this dataset will be introduced first. Then exploratory data analysis will show the overall structure of our dataset, thus providing more information about the dataset. This can be done by checking the mean or variance of the dataset and the overall pattern in the figures. The details of the model specification can be found in section 4.3. In order to compare the proposed model with the preexisting methods (GLM, Univariate model), the analysis result of all three methods will be discussed in section 4.4. To analyze the overall model performance and evaluate the model bias and deviations, a simulation study will be also included at the end of chapter 4.

Finally, some conclusions and future work will be discussed in chapter 5.

Chapter 2

Literature Review

2.1 Generalized Linear Model

In statistics, the simplest way to study the relationship between one or more explanatory variables x_1,...,x_i,...,x_n and the response variable Y is by fitting linear regression model. The basic linear model can only express the expected value of response variable Y_i by a linear combination of unknown constants, which is $Y = \sum_{i=1}^{n}(x_i\beta_i) + e$. The e is the error term which is assumed to follow normal distribution. Using a linear model requires checking normality and linearity. Linear models customarily embody both systematic and random parts, where the errors are usually assumed to be normally distributed. In 1972, the generalized linear model was introduced by Nelder and Wedderburn. (1972). The advantage of generalized linear model is that it can allow the response variables to have random error distribution other than a normal distribution. According to Nelder and Wedderburn. (1972),

the generalized linear model is produced by combining the systematic and random components in the model.

Which generally have three characterizations:

1: The response variable Y are belonging to one of the exponential family of probability distribution.

2: A set of independent exploratory variables X_1,...,X_i,...,X_n and a linear predictor $\eta = X\beta$.

3: A link function g such that $g^{-1}(\eta) = E(Y|X) = \mu$, connecting the mean of the distribution of Y with the link predictor η.

The link function g provides the link between random and systematic components, shows how the expected value of the response Y relates to the linear predictor η of explanatory variables X. The linear predictor η represents a linear combination of parameter β and explanatory variable X. Table 2.1 shows several common distributions with their link function.

Table 2.1: The link function of several common distributions

Distribution	Link Name	Canonical Link $X\beta = g(\mu)$	Inverse Link $\mu = g^{-1}(X\beta)$
Normal	Identity Link	μ	$X\beta$
Poisson	Log Link	$\log \mu$	$\exp(X\beta)$
Gamma	Negative Inverse	$-\frac{1}{\mu}$	$-\frac{1}{X\beta}$
Binomial	Logit Link	$\log \frac{\mu}{1-\mu}$	$\frac{\exp(X\beta)}{1+exp(X\beta)}$
Wald	Inverse Squared	μ^{-2}	$(X\beta)^{-\frac{1}{2}}$

According to Turner (2008), most of the commonly used statistical distributions are members of the exponential family of distributions whose densities can be written in the form:

$$f(y;\theta,\phi) = exp\{\frac{y\theta - b(\theta)}{\phi + c(y,\phi)}\} \tag{2.1}$$

where ϕ is the dispersion parameter and θ is the canonical parameter. It

can be shown that

$$E(Y) = b^{'}(\theta) = \mu \tag{2.2}$$

and

$$Var(Y) = \phi b^{''}(\theta) = \phi V(\mu) \tag{2.3}$$

A single algorithm can be used to estimate the parameters of an exponential family glm using maximum likelihood. The log-likelihood for the sample $y_1, y_2, ..., y_n$ is:

$$l = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i, \phi_i) \tag{2.4}$$

With single dispersion parameter ϕ and known prior weights a_i, we assumed that $\phi_i = \frac{\phi}{a_i}$, then by solving the score equation, the maximum likelihood estimates are:

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{a_i(y_i - \mu_i)}{V(\mu_i)} \times \frac{x_{ij}}{g^{'}(\mu_i)} \tag{2.5}$$

A general method of solving score equations is the iterative algorithm Fisher's Method of Scoring. In the r^{th} iteration, the new estimate $\beta^{(r+1)}$ is obtained from the previous estimates $\beta^{(r)}$ by:

$$\beta^{(r+1)} = \beta^{(r)} + s(\beta^{(r)})E(H(\beta^{(r)}))^{-1} \tag{2.6}$$

where H is the matrix of second derivatives of the log-likelihood. It turns out that the updates can be written as:

$$\beta^{(r+1)} = (X^T W^{(r)} X)^{-1} X^T W^{(r)} Z^{(r)} \tag{2.7}$$

where weights $W^{(r)} = diag(w_i)$ Hence the estimates can be found using an IRWLS algorithm:

1: Start with initial estimates $\mu_i^{(r)}$

2: Calculate working responses Z^r and working weights $w^{(r)}$

3: Calculate $\beta^{(r+1)}$ by weighted least squares

4: Repeat 2 and 3 till convergence

2.2 Tweedie Exponential Dispersion Model

In statistics, Tweedie distribution is a special case of exponential dispersion models and a newcomer to the generalized linear model framework. The exponential dispersion model is a set of probability distributions that represents a generalisation of the natural exponential family. An important advantage of exponential dispersion models is the elegant asymptotic theory, which generalizes the analysis of deviance for generalized linear models, giving rise to asymptotic versions of the familiar t-test, F-test, and X^2-tests from linear normal theory (Jørgensen, 1987). In general, exponential dispersion models, which are linear exponential families with a dispersion parameter, are the prototype response distributions for generalized linear models. The Tweedie family comprises those exponential dispersion models with power mean-variance relationships (Dunn, P.K. et al, 2005). According to Jørgensen (1987), an exponential dispersion model is characterized by its

variance function $V(\mu)$, which corresponds to power variance functions:

$$V(\mu) = \mu^p \tag{2.8}$$

where V is the variance function and the different values of the power index parameter p indicates the different kinds of distributions. For example, the power index parameter p is set to 0 for normal distribution, set to 1 for Poisson distribution. More details about the power index parameter can be found in Table 2.2.

Because Tweedie distribution is a special case of exponential dispersion models, so many exponential dispersion models have variance functions that are asymptotically of the Tweedie form (Ma, 1999). For Tweedie exponential dispersion model, the power variance function can be expressed as:

$$V(Y) = \sigma^2\mu^p \tag{2.9}$$

where σ^2 is the dispersion parameter and the condition is that Y follows the Tweedie distribution. This can be denoted by $Tw_p(\mu, \sigma^2)$, if $E(Y) = \mu$ and $Var(Y) = \sigma^2\mu^p$. In addition, Y follows a Tweedie exponential dispersion distribution $Tw_p(\mu, \sigma^2)$, if its density is of the form:

$$f_p(y;\mu.\sigma^2)\begin{cases} c_p(y;\sigma^2)\exp\{\frac{1}{\sigma^2}(\frac{y\mu^{1-p}}{1-p}-\frac{\mu^{2-p}}{2-p})\}, & \text{if } p\neq 1,2,\\ c_2(y;\sigma^2)\exp[-\frac{1}{\sigma^2}\{\frac{y}{\mu}+\log(\mu)\}], & \text{if } p=2,\\ c_1(y)\exp\{y\log(\mu)-\mu\}, & \text{if } p=1 \end{cases}$$

The detail of normalizing constant $c_p(y;\sigma^2)$ can be found in (Jørgensen, 1987). The more details about the power index p can be found in Table 2.2.

Table 2.2: Power index parameters p for different distribution

Distribution	Power Index Parameter p
Normal	$p=0$
Poisson	$p=1$
Compound Poisson	$1<p<2$
Gamma	$p=2$
Wald	$p=3$

2.3 Previous Work

Bekele and Shen (2005) developed a Bayesian approach to jointly modelling a binary toxicity outcome and a continuous biomarker expression outcome. More recently, Lee et al. (2015) have published a design for adaptively and dynamically optimizing the patient's dose which is the covariate in each of two cycles based on binary outcomes and continuous outcomes measured at a single time point. Moreover, based on the previous work done by Bekele

and Shen (2005), Ezzalfani et al. (2018) have developed joint modelling of a binary and a continuous outcome. Their main goal was trying to find the optimal dose of targeted treatment in oncology. Because of the drugs appear likely to induce toxicity after the first cycle, so they measured both binary toxicity and efficacy outcomes at another cycle. Therefore, they were modelling jointly a continuous biomarker of activity and a binary toxicity variable. Latent normal variable and probit functions are a versatile means to model jointly two repeated variables that can be either two categorical variables or continuous variables. So, they also introduce a continuous latent variable to model the continuous efficacy variables and the binary toxicity responses jointly as a function of dose. Maximum likelihood estimators are used to estimate the parameters of the joint model as well.

Let i index the n patients, and x_k index the dose that is administered to patient i at cycle j, among k possible doses. The pair of outcomes of patient i at the jth cycle is denoted by (T_{ij}, Y_{ij}), with T_{ij} the binary toxicity DLT and Y_{ij} the continuous biomarker. By using the probit model, with the assumption that there is a latent variable Z_{ij}, normally distributed with variance equal to 1, which is related to the binary observed toxicity as:

$$T_{ij} = \begin{cases} 0, & \text{if } Z_{ij} \leq 0 \\ 1, & \text{if } Z_{ij} > 0 \end{cases} \tag{2.10}$$

$j = 1, 2$ and $i = 1, 2, ...,$ n

For each cycle, a linear model between the latent variable Z_{ij} and the dose x_k is assumed:

$$Z_{ij} = \alpha_j + \beta_j x_k + \epsilon_{ij} \tag{2.11}$$

where $\alpha_j \in R$, $\beta_j \in R$ and $\epsilon_j \sim N(0,1)$

For each cycle, the probability of toxicity can be expressed as

$$Pr(T_{ij} = 1|x_k) = Pr(Z_{ij} > 0|x_k) = \phi(\alpha_j + \beta_j x_k) \tag{2.12}$$

where ϕ is the cumulative density function of a standard Gaussian variable. The relationship between the biomarker Y_{ij} and the dose x_k is described by using the regression model

$$Y_{ij} = \alpha_j^{'} + \beta_j^{'} x_k + \epsilon_{ij}^{'}, \tag{2.13}$$

$\epsilon_{ij}^{'} \sim N(0, \sigma_Y^2)$ and $k \in 1, ..., K$. The continuous latent variable is used to model jointly the continuous biomarker outcome and the binary toxicity outcome by using the joint distribution of the four Gaussian variables (Z_{i1},Y_{i1},Z_{i2},Y_{i2}). The density function of (Z_{i1},Y_{i1},Z_{i2},Y_{i2}) is then a multivariate Gaussian density of dimension 4, which can be written as

$$f_{Z_1,Y_1,Z_2,Y_2(Z_{i1},Y_{i1},Z_{i2},Y_{i2})} = \frac{1}{(2\Pi)^2|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(V-\mu)\Sigma^{-1}(V-\mu)\} \tag{2.14}$$

where $V = (Z_{i1}, Y_{i1}, Z_{i2}, Y_{i2})^{'}$, $\mu = (\mu_{Z1,Y1,Z2,Y2})^{'}$, the mean vector, and Σ is the variance-covariance matrix of (Z, Y). The correlation between the binary toxicity and continuous biomarker outcome of cycles 1 and 2 can be described by the variance-covariance matrix.

2.4 Quasi Likelihood Function

The concept of a dispersion parameter is also used in the method of quasi likelihood. The quasi likelihood technique is used for estimating regression coefficients without fully specifying the distribution of the observed data. Let the observed vector of random variables Y be of length N and suppose that the log likelihood considered initially as a function of the N-dimensional parameter θ and an additional parameter σ^2 may be written in the form

$$\sigma^2\{y^T\theta - b(\theta) - c(y, \sigma)\} \tag{2.15}$$

for suitably chosen functions $b(\theta)$ and $c(y, \sigma)$. According to McCullagh (1983), the systematic component may be expressed in terms of a regression equation

$$E(Y) = \mu = \mu(\beta) \tag{2.16}$$

and the generalized least squares equations for the parameters in 2.4 can be written as

$$D^{'}V^{-1}\{y - \mu(\hat{\beta})\} = 0 \tag{2.17}$$

where $D = d\mu/d\beta$ and V^{-1} is a generalized inverse of variance function V. Similarly, for independent observation $Y_1, ..., Y_i, ...Y_n$ with $Var[Y_i] = \phi V(\mu_i)$ the quasi score function can be written as

$$U(\beta; y) = D^{'} V^{-1}(y - \mu)/\phi \tag{2.18}$$

where $D = d\mu/d\beta$ and V^{-1} is a generalized inverse of variance function V.

Chapter 3

A Joint Mixed Model for Clustered Binary and Continuous Outcomes

3.1 Model Specification

Let Y_{ij1} represents the binary response recorded on the jth ($j = 1, 2, ..., N_i$) subject within the ith ($i = 1, 2, ..., m$) cluster. Let Y_{ij2} represents the continuous response recorded on the jth ($j = 1, 2, ..., N_i$) subject within the ith ($i = 1, 2, ..., m$) cluster. Let $Y_i = (Y_{i11}, Y_{i21}, ..., Y_{ij1}, ..., Y_{iN_i1}, Y_{i12}, ..., Y_{ij2}, ..., Y_{iN_i2})^T$ represents the vector of the responses. The responses are assumed to be Bernoulli distributed or Tweedie distributed depending on the response type. Let $U = (U_1, ..., U_i, ..., U_m)^T$ represents the vector of the first level random effects. Let $V = (V_{11}, ..., V_{1j}, ..., V_{1N_1}, V_{21}, ..., V_{mN_m})^T$ and

$W = (W_{11}, ..., W_{1j}, ..., W_{1N_1}, W_{21}, ..., W_{mN_m})^T$ represents the vector of the two second level random effects, where the cluster-specific random effects U_i for the ith cluster and subject random effect V_{ij} and W_{ij} for the response from the two response levels on the jth subject within ith cluster. Three assumptions are used in our model.

3.1.1 Assumption 1

The first level random effects $U_1, ..., U_i, ..., U_m$ are cluster-specific random effects. U_i distributed on $(0, 1)$. They are independently and identically distributed with

$$E(U_i) = \frac{1}{2} \quad \text{and} \quad \text{Var}(U_i) = \sigma^2 \tag{3.1}$$

The N_i is the total number of subjects in the i^{th} cluster.

3.1.2 Assumption 2

Because there are two types of response variables, which are binary response and continuous response. Therefore, two second level random effects V_{ij} and W_{ij} are required in this model.

The second level random effects V_{ij} are subject-specific random effects, conditionally independent given the cluster-specific random effects U_i and corresponding to the binary response. If there are multiple measurements on subject j within cluster i, then they are positive and can be expressed as

$$E(V_{ij}|U) = U_i \quad \text{and} \quad \text{Var}(V_{ij}|U) = \eta^2 U_i \tag{3.2}$$

If there is only one measurement on subject j within cluster i, then they are positive and can be expressed as

$$E(V_{ij}|U) = U_i \quad \text{and} \quad \text{Var}(V_{ij}|U) = 0 \tag{3.3}$$

which is equivalent to

$$V_{ij} = U_i \tag{3.4}$$

The second level random effects W_{ij} are subject-specific random effects, conditionally independent given the cluster-specific random effects U_i and corresponding to the continuous response. They are positive and can be expressed as

$$E(W_{ij}|U) = U_i \quad \text{and} \quad \text{Var}(W_{ij}|U) = \tau^2 U_i, \quad W_{ij} > 0 \tag{3.5}$$

All second level random effects are conditionally independent on the first level random effect.

3.1.3 Assumption 3

The third assumption is the components of Y are conditionally independent given the random effects Z, where Z is $Z = (U, V, W)$. The conditional distribution of Y_{ij1} and Y_{ij2}, given Z, depends on random effects V_{ij} and W_{ij}, which are

$$Y_{ij1}|Z \sim Ber(\pi_{ij} V_{ij}) \tag{3.6}$$

$$Y_{ij2}|Z \sim Tw_p(\mu_{ij}W_{ij}, \epsilon^2 W_{ij}^{1-p}) \tag{3.7}$$

where $log(\pi_{ij}/(1-\pi_{ij})) = logit(\pi_{ij}) = x_{ij}^{'}\gamma$ with vector of covariates x_{ij}, measured at the binary response on the jth subject within ith cluster and regression parameter vector γ. And $\mu_{ij} = \exp(x_{ij}^{'}\beta)$ with vector of covariates x_{ij}, measured at the continuous response on the jth subject within ith cluster and regression parameter vector β. In (3.6), Ber is the Bernoulli family. In (3.7), Tw_p is the Tweedie exponential family with index p.

$$E(Y_{ij1}|Z) = V_{ij}\pi_{ij} \tag{3.8}$$

$$\text{Var}(Y_{ij1}|Z) = (V_{ij}\pi_{ij})(1 - V_{ij}\pi_{ij}) = V_{ij}\pi_{ij} - V_{ij}^2\pi_{ij}^2 \tag{3.9}$$

$$E(Y_{ij2}|Z) = \mu_{ij}W_{ij} \tag{3.10}$$

$$\text{Var}(Y_{ij2}|Z) = \epsilon^2 W_{ij}^{(1-p)}(\mu_{ij}W_{ij})^p = \epsilon^2 \mu_{ij}^p W_{ij} \tag{3.11}$$

With the expected value of random variable $Y_{ij2}|Z$ which is showed in (3.10) and the variance of random variable $Y_{ij2}|Z$ which is showed in (3.11), this Tweedie family are often called as the power-variance family.

3.2 Moment Structure

For the binary response level, the unconditional expectation and variance of the response Y_{ij1} can be expressed as:

$$E(Y_{ij1}) = E(E(Y_{ij1}|Z))$$

$$\mathrm{Var}(Y_{ij1}) = E(\mathrm{Var}(Y_{ij1}|Z)) + \mathrm{Var}(E(Y_{ij1}|Z))$$

With (3.1), (3.3) and (3.6), the unconditional expected value of the Y_{ij1} is:

$$\begin{aligned} E(Y_{ij1}) &= E(E(Y_{ij1}|Z)) \\ &= E(\pi_{ij}V_{ij}) \\ &= \pi_{ij}E(V_{ij}) \\ &= \pi_{ij}E(U_i) \\ &= \frac{1}{2}\pi_{ij} \end{aligned} \tag{3.12}$$

And the unconditional variance of the Y_{ij1} is:

$$\begin{aligned} \mathrm{Var}(Y_{ij1}) &= E(\mathrm{Var}(Y_{ij1}|Z)) + \mathrm{Var}(E(Y_{ij1}|Z)) \\ &= E[V_{ij}\pi_{ij} - V_{ij}^2\pi_{ij}^2] + \mathrm{Var}(V_{ij}\pi_{ij}) \\ &= \pi_{ij}E(V_{ij}) - \pi_{ij}^2E(V_{ij}^2) + \pi_{ij}^2\mathrm{Var}(V_{ij}) \\ &= \pi_{ij}E[E(V_{ij}|U)] - \pi_{ij}^2[\mathrm{Var}(V_{ij}) + E(V_{ij})^2] + \pi_{ij}^2\mathrm{Var}(V_{ij}) \\ &= \pi_{ij}E(U_i) - \pi_{ij}^2E(V_{ij})^2 \\ &= \frac{1}{2}\pi_{ij} - \pi_{ij}^2(\frac{1}{2})^2 \\ &= \frac{1}{2}\pi_{ij}(1 - \frac{1}{2}\pi_{ij}) \end{aligned} \tag{3.13}$$

The unconditional covariance of Y_{ij1} and $Y_{i'j'1}$ can be derived as below:

$$\begin{aligned}
\mathrm{Cov}(Y_{ij1}, Y_{i'j'1}) &= E(\mathrm{Cov}(Y_{ij1}, Y_{i'j'1}|Z)) + \mathrm{Cov}(E(Y_{ij1}|Z), E(Y_{i'j'1}|Z)) \\
&= \delta(i,i')\delta(j,j')E(V_{ij}\pi_{ij} - V_{ij}^2\pi_{ij}^2) + \mathrm{Cov}(V_{ij}\pi_{ij}, V_{i'j'}\pi_{i'j'}) \\
&= \delta(i,i')\delta(j,j')[\frac{1}{2}\pi_{ij} - \pi_{ij}^2(\mathrm{Var}(V_{ij}) + E(V_{ij})^2)] + \\
&\quad \pi_{ij}\pi_{i'j'}\mathrm{Cov}(V_{ij}, V_{i'j'}) \\
&= \delta(i,i')\delta(j,j')[\frac{1}{2}\pi_{ij} - \pi_{ij}^2(0 + \mathrm{Var}(U_i) + E(U_i)^2)] + \\
&\quad \pi_{ij}\pi_{i'j'}\mathrm{Cov}(V_{ij}, V_{i'j'}) \\
&= \delta(i,i')\delta(j,j')[\frac{1}{2}\pi_{ij} - \pi_{ij}^2(\mathrm{Var}(U_i) + E(U_i)^2)] + \\
&\quad \pi_{ij}\pi_{i'j'}[E(\mathrm{Cov}(V_{ij}, V_{i'j'}|U)) + \mathrm{Cov}(E(V_{ij}|U), E(V_{i'j'}|U))] \\
&= \delta(i,i')\delta(j,j')[\frac{1}{2}\pi_{ij} - \pi_{ij}^2(\mathrm{Var}(U_i) + E(U_i)^2)] + \\
&\quad \pi_{ij}\pi_{i'j'}[\delta(i,i')\sigma^2] \\
&= \frac{1}{2}\delta(i,i')\delta(j,j')[\pi_{ij} - \pi_{ij}^2(2\sigma^2 + \frac{1}{2})] + \delta(i,i')\pi_{ij}\pi_{i'j'}\sigma^2
\end{aligned} \tag{3.14}$$

For the continuous response level, The unconditional expectation and variance of the response Y_{ij2} can be expressed as:

$$E(Y_{ij2}) = E(E(Y_{ij2}|Z))$$

$$\mathrm{Var}(Y_{ij2}) = E(\mathrm{Var}(Y_{ij2}|Z)) + \mathrm{Var}(E(Y_{ij2}|Z))$$

With (3.1), (3.5) and (3.7), the unconditional expected value of the Y_{ij2} is:

$$\begin{aligned} E(Y_{ij2}) &= E(E(Y_{ij2}|Z)) \\ &= E(\mu_{ij}W_{ij}) \\ &= \mu_{ij}E(W_{ij}) \\ &= \mu_{ij}E(U_i) \\ &= \frac{1}{2}\mu_{ij} \end{aligned} \tag{3.15}$$

And the unconditional variance of the Y_{ij2} is:

$$\begin{aligned} \mathrm{Var}(Y_{ij2}) &= E(\mathrm{Var}(Y_{ij2}|Z)) + \mathrm{Var}(E(Y_{ij2}|Z)) \\ &= E(\epsilon^2\mu_{ij}^p W_{ij}) + \mathrm{Var}(\mu_{ij}W_{ij}) \\ &= \epsilon^2\mu_{ij}^p E(W_{ij}) + \mu_{ij}^2\mathrm{Var}(W_{ij}) \\ &= \epsilon^2\mu_{ij}^p E[E(W_{ij}|U)] + \mu_{ij}^2[E(\mathrm{Var}(W_{ij}|U)) + \mathrm{Var}(E(W_{ij}|U))] \\ &= \epsilon^2\mu_{ij}^p E(U_i) + \mu_{ij}^2[E(\tau^2 U_i) + \mathrm{Var}(U_i)] \\ &= \frac{1}{2}\epsilon^2\mu_{ij}^p + \mu_{ij}^2[\tau^2 E(U_i) + \sigma^2] \\ &= \frac{1}{2}\epsilon^2\mu_{ij}^p + \mu_{ij}^2(\frac{1}{2}\tau^2 + \sigma^2) \end{aligned} \tag{3.16}$$

The unconditional covariance of Y_{ij2} and $Y_{i'j'2}$ can be derived as below:

$$\begin{aligned}
\mathrm{Cov}(Y_{ij2}, Y_{i'j'2}) &= E(\mathrm{Cov}(Y_{ij2}, Y_{i'j'2}|Z)) + \mathrm{Cov}(E(Y_{ij2}|Z), E(Y_{i'j'2}|Z)) \\
&= E(\delta(i,i')\delta(j,j')\epsilon^2\mu_{ij}^p W_{ij}) + \mathrm{Cov}(\mu_{ij}W_{ij}, \mu_{i'j'}W_{i'j'}) \\
&= \delta(i,i')\delta(j,j')\epsilon^2\mu_{ij}^p E(W_{ij}) + \mu_{ij}\mu_{i'j'}\mathrm{Cov}(W_{ij}, W_{i'j'}) \\
&= \delta(i,i')\delta(j,j')\epsilon^2\mu_{ij}^p E(W_{ij}) + \\
&\quad \mu_{ij}\mu_{i'j'}[E(\mathrm{Cov}(W_{ij}, W_{i'j'}|U)) + \mathrm{Cov}(E(W_{ij}|U), E(W_{i'j'}|U))] \\
&= \frac{1}{2}\delta(i,i')\delta(j,j')\epsilon^2\mu_{ij}^p + \\
&\quad \mu_{ij}\mu_{i'j'}[\delta(i,i')\delta(j,j')\tau^2 E(U_i) + \delta(i,i')\sigma^2] \\
&= \frac{1}{2}\delta(i,i')\delta(j,j')(\epsilon^2\mu_{ij}^p + \mu_{ij}\mu_{i'j'}\tau^2) + \delta(i,i')\mu_{ij}\mu_{i'j'}\sigma^2
\end{aligned} \tag{3.17}$$

In addition, the unconditional covariance of Y_{ij1} and $Y_{i'j'2}$ can be derived as below:

$$\begin{aligned}
\mathrm{Cov}(Y_{ij1}, Y_{i'j'2}) &= E(\mathrm{Cov}(Y_{ij1}, Y_{i'j'2}|Z)) + \mathrm{Cov}(E(Y_{ij1}|Z), E(Y_{i'j'2}|Z)) \\
&= 0 + \mathrm{Cov}(V_{ij}\pi_{ij}, \mu_{i'j'}W_{i'j'}) \\
&= \delta(i,i')\pi_{ij}\mu_{i'j'}\mathrm{Cov}(V_{ij}, W_{i'j'}) \\
&= \delta(i,i')\pi_{ij}\mu_{i'j'}[E(\mathrm{Cov}(V_{ij}, W_{i'j'}|Z)] + \\
&\quad \mathrm{Cov}[E(V_{ij}|Z), E(W_{i'j'}|Z)]] \\
&= \delta(i,i')\pi_{ij}\mu_{i'j'}[0 + \mathrm{Var}(U_i)] \\
&= \delta(i,i')\pi_{ij}\mu_{i'j'}\sigma^2
\end{aligned} \tag{3.18}$$

while,

$$\delta(i,i') = \begin{cases} 1, & i = i' \\ 0, & i \neq i' \end{cases} \tag{3.19}$$

and

$$\delta(j,j') = \begin{cases} 1, & j = j' \\ 0, & j \neq j' \end{cases} \tag{3.20}$$

Hence, there are six cases and need to be discussed and calculated separately.

Case 1: Same cluster, same subject and with first response level

If $i = i', j = j'$ and with first response level:

$$\begin{aligned} \text{Cov}(Y_{ij1}, Y_{ij1}) &= \text{Var}(Y_{ij1}) \\ &= \frac{1}{2}\pi_{ij}(1 - \frac{1}{2}\pi_{ij}) \end{aligned} \tag{3.21}$$

Case 2: Same cluster, same subject and with second response level

If $i = i', j = j'$ and with second response level:

$$\begin{aligned} \text{Cov}(Y_{ij2}, Y_{ij2}) &= \text{Var}(Y_{ij2}) \\ &= \frac{1}{2}\epsilon^2\mu_{ij}^p + \mu_{ij}^2(\frac{1}{2}\tau^2 + \sigma^2) \end{aligned} \tag{3.22}$$

Case 3: Same cluster, different subjects and with first response level

If $i = i', j \neq j'$ and with first response level:

$$\begin{aligned} \text{Cov}(Y_{ij1}, Y_{ij'1}) &= E(\text{Cov}(Y_{ij1}, Y_{ij'1}|Z)) + \text{Cov}(E(Y_{ij1}|Z), E(Y_{ij'1}|Z)) \\ &= \pi_{ij}\pi_{ij'}\sigma^2 \end{aligned} \tag{3.23}$$

Case 4: Same cluster, different subjects and with second response level

If $i = i', j \neq j'$ and with second response level:

$$\begin{aligned} \text{Cov}(Y_{ij2}, Y_{ij'2}) &= E(\text{Cov}(Y_{ij2}, Y_{ij'2}|Z)) + \text{Cov}(E(Y_{ij2}|Z), E(Y_{ij'2}|Z)) \\ &= \mu_{ij}\mu_{ij'}\sigma^2 \end{aligned} \tag{3.24}$$

Case 5: Same cluster, different subjects and with different response levels

If $i = i', j \neq j'$ and with different response levels:

$$\begin{aligned} \text{Cov}(Y_{ij1}, Y_{ij'2}) &= E(\text{Cov}(Y_{ij1}, Y_{ij'2}|Z)) + \text{Cov}(E(Y_{ij1}|Z), E(Y_{ij'2}|Z)) \\ &= \pi_{ij}\mu_{ij'}\sigma^2 \end{aligned} \tag{3.25}$$

Case 6: Otherwise

$$\text{Cov}(Y_{ijk}, Y_{i'j'k'}) = 0 \tag{3.26}$$

In general, the covaraince of Y can be written in matrix form as below:

$$\mathrm{Cov}(Y) = \begin{bmatrix} \mathrm{Cov}(Y_1) & 0 & ... & 0 \\ 0 & \mathrm{Cov}(Y_2) & ... & 0 \\ ... & ... & \mathrm{Cov}(Y_i) & ... \\ 0 & 0 & ... & \mathrm{Cov}(Y_m) \end{bmatrix} \tag{3.27}$$

where $\mathrm{Cov}(Y_i)$ are:

$$\mathrm{Cov}(Y_i) = \begin{bmatrix} \mathrm{Var}(Y_{i1}) & \mathrm{Cov}(Y_{i1}, Y_{i2}) \\ \mathrm{Cov}(Y_{i2}, Y_{i1}) & \mathrm{Var}(Y_{i2}) \end{bmatrix} \tag{3.28}$$

where $\mathrm{Cov}(Y_{ik}, Y_{ik})$ are:

$$\mathrm{Cov}(Y_{ik}, Y_{ik}) = \begin{bmatrix} \mathrm{Var}(Y_{i1k}) & \mathrm{Cov}(Y_{i1k}, Y_{i2k}) & ... & \mathrm{Cov}(Y_{i1k}, Y_{iN_ik}) \\ \mathrm{Cov}(Y_{i2k}, Y_{i1k}) & \mathrm{Var}(Y_{i2k}) & ... & \mathrm{Cov}(Y_{i2k}, Y_{iN_2k}) \\ ... & ... & \mathrm{Var}(Y_{ijk}) & ... \\ \mathrm{Cov}(Y_{iN_ik}, Y_{i1k}) & \mathrm{Cov}(Y_{iN_ik}, Y_{i2k}) & ... & \mathrm{Var}(Y_{iN_ik}) \end{bmatrix} \tag{3.29}$$

3.3 The best linear unbiased predictors of random effects

The best linear unbiased predictors of random effects of the cluster-specific random effects, U given Y can be predicted as:

$$\hat{U} = E(U) + \text{Cov}(U, Y)\text{Var}^{-1}(Y)(Y - E(Y)) \tag{3.30}$$

where $E(U) = (\frac{1}{2}, ..., \frac{1}{2}, ...\frac{1}{2})^T$ of dimension $m \times 1$. The covaraince structure between U and first response level Y_1 can be derived as:

$$\begin{aligned} \text{Cov}(U_i, Y_{ij1}) &= E(\text{Cov}(U, Y|Z)) + \text{Cov}(E(U|Z), E(Y|Z)) \\ &= \pi_{ij}\sigma^2 \end{aligned} \tag{3.31}$$

And the covaraince structure between U and second response level Y_2 can be derived as:

$$\begin{aligned} \text{Cov}(U_i, Y_{ij2}) &= E(\text{Cov}(U, Y|Z)) + \text{Cov}(E(U|Z), E(Y|Z)) \\ &= \mu_{ij}\sigma^2 \end{aligned} \tag{3.32}$$

With (3.31) and (3.32), the covariance structure between U and Y can be obtained as:

$$\mathrm{Cov}(U,Y) = \begin{bmatrix} \mathrm{Cov}(U_1,Y_1) & 0 & ... & 0 \\ 0 & \mathrm{Cov}(U_2,Y_2) & ... & 0 \\ ... & ... & \mathrm{Cov}(U_i,Y_i) & ... \\ 0 & 0 & ... & \mathrm{Cov}(U_m,Y_m) \end{bmatrix} \tag{3.33}$$

where:

$$\mathrm{Cov}(U_i,Y_i) = (\pi_{i1}\sigma^2, \pi_{i2}\sigma^2, ..., \pi_{iN_i}\sigma^2, \mu_{i1}\sigma^2, ..., \mu_{iN_i}\sigma^2) \tag{3.34}$$

The best linear unbiased predictors of random effects of the subject-specific random effects, V given binary response level Y_1 can be predicted as:

$$\hat{V} = E(V) + \mathrm{Cov}(V,Y)\mathrm{Var}^{-1}(Y)(Y - E(Y)) \tag{3.35}$$

where $E(V) = (1/2, 1/2, ..., 1/2)'$ of dimension $N_i * N_i$ and $\mathrm{Cov}(V_i, Y_{i1})$ is a $N_i * N_i$ matrix can be calculated as:

$$\begin{aligned} \mathrm{Cov}(V_i,Y_{i1}) &= E(\mathrm{Cov}(V_i,Y_{i1}|Z)) + \mathrm{Cov}(V_i,E(Y_{i1}|Z)) \\ &= 0 + \mathrm{Cov}(V_i,\pi_i V_i) \\ &= \mathrm{Var}(V_i)\mathrm{diag}(\pi_i') \end{aligned} \tag{3.36}$$

where $\mathrm{Var}(V_i)$ can be expressed as:

$$\mathrm{Var}(V_i) = \sigma^2 \mathbf{1}_{N_i} \mathbf{1}_{N_i}^{'} \tag{3.37}$$

or can be written in matrix form as:

$$\mathrm{Var}(V_i) = \begin{bmatrix} \sigma^2 & \sigma^2 & ... & \sigma^2 \\ \sigma^2 & \sigma^2 & ... & \sigma^2 \\ ... & ... & ... & ... \\ \sigma^2 & \sigma^2 & ... & \sigma^2 \end{bmatrix} \tag{3.38}$$

The $\mathbf{1}_{N_i}$ denotes a $N_i \times 1$ matrix of ones and $\mathbf{I}_{N_i}$ is a identity $N_i \times N_i$ matrix.

And $\mathrm{diag}(\pi_i)$ is defined as:

$$\mathrm{diag}(\pi_i) = \begin{bmatrix} \pi_{i1} & 0 & ... & 0 \\ 0 & \pi_{i2} & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \pi_{iN_i} \end{bmatrix} \tag{3.39}$$

Then, the covariance between V_i and Y_{i1} in matrix expression can be expressed as:

$$\mathrm{Cov}(V_i, Y_{i1}) = (\sigma^2 \mathbf{1}_{N_i} \mathbf{1}_{N_i}^{'}) \mathrm{diag}(\pi_i) \tag{3.40}$$

Also, the covariance between V_i and Y_{i2} in matrix expression can be ex-

pressed as

$$\begin{aligned}\mathrm{Cov}(V_i, Y_{i2}) &= E(\mathrm{Cov}(V_i, Y_{i2}|Z)) + \mathrm{Cov}(V_i, E(Y_{i2}|Z)) \\ &= 0 + \mathrm{Cov}(V_i, \mu_i W_i) \\ &= \mathrm{Cov}(V_i, W_i)\mu_i^{'} \\ &= \sigma^2 \mathbf{1}_{N_i} \mathbf{1}_{N_i}^{'} \mathrm{diag}(\mu_i)\end{aligned} \tag{3.41}$$

Therefore, the covariance between V_i and Y_i can be expressed as:

$$\mathrm{Cov}(V_i, Y_i) = \begin{bmatrix} \mathrm{Cov}(V_i, Y_{i1}) & \mathrm{Cov}(V_i, Y_{i2}) \end{bmatrix} \tag{3.42}$$

The best linear unbiased predictors of random effects of the subject-specific random effects, W given continuous response level Y_2 can be predicted as:

$$\hat{W} = E(W) + \mathrm{Cov}(W, Y)\mathrm{Var}^{-1}(Y)(Y - E(Y)) \tag{3.43}$$

where $E(W) = (1/2, 1/2, ..., 1/2)^{'}$ of dimension $N_i * N_i$ and $\mathrm{Cov}(W_i, Y_{i2})$ is a $N_i * N_i$ matrix can be calculated as:

$$\begin{aligned}\mathrm{Cov}(W_i, Y_{i2}) &= E(\mathrm{Cov}(W_i, Y_{i2}|Z)) + \mathrm{Cov}(W_i, E(Y_{i2}|Z)) \\ &= 0 + \mathrm{Cov}(W_i, \mu_i W_i) \\ &= \mathrm{Var}(W_i)\mathrm{diag}(\mu_i')\end{aligned} \tag{3.44}$$

where $\mathrm{Var}(W_i)$ can be expressed as:

$$\mathrm{Var}(W_i) = \sigma^2 \mathbf{1}_{N_i} \mathbf{1}_{N_i}^{'} + \frac{1}{2}\tau^2 \mathbf{I}_{N_i} \tag{3.45}$$

or can be written in matrix form as:

$$\mathrm{Var}(W_i) = \begin{bmatrix} \frac{1}{2}\tau^2 + \sigma^2 & \sigma^2 & ... & \sigma^2 \\ \sigma^2 & \frac{1}{2}\tau^2 + \sigma^2 & ... & \sigma^2 \\ ... & ... & ... & ... \\ \sigma^2 & \sigma^2 & ... & \frac{1}{2}\tau^2 + \sigma^2 \end{bmatrix} \tag{3.46}$$

And $\mathrm{diag}(\mu_i)$ is defined as:

$$diag(\mu_i) = \begin{bmatrix} \mu_{i1} & 0 & ... & 0 \\ 0 & \mu_{i2} & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \mu_{iN_i} \end{bmatrix} \tag{3.47}$$

Then, the covariance between W_i and Y_{i2} in matrix expression can be expressed as:

$$\mathrm{Cov}(W_i, Y_{i2}) = (\sigma^2 \mathbf{1}_{N_i} \mathbf{1}_{N_i}^{'} + \frac{1}{2}\tau^2 \mathbf{I}_{N_i})\mathrm{diag}(\mu_i) \tag{3.48}$$

Also, the covariance between W_i and Y_{i1} in matrix expression can be ex-

pressed as

$$\begin{aligned}\mathrm{Cov}(W_i, Y_{i1}) &= E(\mathrm{Cov}(W_i, Y_{i1}|Z)) + \mathrm{Cov}(W_i, E(Y_{i1}|Z)) \\ &= 0 + \mathrm{Cov}(W_i, \pi_i W_i) \\ &= \mathrm{Cov}(W_i, V_i)\pi_i^{'} \\ &= \sigma^2 \mathbf{1}_{N_i} \mathbf{1}_{N_i}^{'} \mathrm{diag}(\pi_i)\end{aligned} \tag{3.49}$$

Therefore, the covariance between W_i and Y_i can be expressed as:

$$\mathrm{Cov}(W_i, Y_i) = \begin{bmatrix}\mathrm{Cov}(W_i, Y_{i1}) & \mathrm{Cov}(W_i, Y_{i2})\end{bmatrix} \tag{3.50}$$

The covariance between two random effects V_i and W_i can be calculated as:

$$\begin{aligned}\mathrm{Cov}(V_i, W_i) &= E[\mathrm{Cov}(V_i, W_i|U)] + \mathrm{Cov}[E(V_i|U), E(W_i|U)] \\ &= 0 + \mathrm{Cov}(U_i, U_i) \\ &= \sigma^2\end{aligned} \tag{3.51}$$

The covariance between two random effects V_i and W_i in matrix form can be expressed as:

$$\mathrm{Cov}(V_i, W_i) = \begin{bmatrix}\sigma^2 & \sigma^2 & ... & \sigma^2 \\ \sigma^2 & \sigma^2 & ... & \sigma^2 \\ ... & ... & ... & ... \\ \sigma^2 & \sigma^2 & ... & \sigma^2\end{bmatrix} \tag{3.52}$$

which can be simplified as:

$$\text{Cov}(V_i, W_i) = \sigma^2 \mathbf{1}_m \mathbf{1}_m^{'} \tag{3.53}$$

3.4 Estimation of Parameters

In this section, we will discuss the estimation of the regression parameters, γ and β when the dispersion parameters are known. In addition, the unknown random effect parameters σ^2, τ^2 and ϵ^2 will be discussed.

3.4.1 Estimation of Regression Parameters

According to Ma (2007), the estimation for the regression parameters under the assumption that the dispersion parameters are known can be calculated by differentiate the partially observed 'joint' log-likelihood of the Tweedie mixed model for the data and random effects with respect to β which generate the partially observed 'joint' score function. After replacing the random effects with their BLUP predictors, we get an unbiased estimating function for the regression parameters β and can be expressed as:

$$\psi(\beta) = \sum_{i=1}^{m} \sum_{j=1}^{N_i} X_{ij2}^{'} \frac{\mu_{ij}^{1-p}(\beta)}{\epsilon^2} [y_{ij2} - \hat{W}_{ij2}(\beta)\mu_{ij2}(\beta)] \tag{3.54}$$

In addition, based on Ma (2007), the solution to the equation $\psi(\beta) = 0$ gives estimates of β which are consistent and asymptotically normal with asymptotic mean β and asymptotic variance given by the inverse of the

sensitivity matrix $S(\boldsymbol{\beta}) = E_{\boldsymbol{\beta}}\{\partial\psi(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\}$.

However, Tweedie cannot handle binary cases. Due to our first response are binary variables, we cannot use the same method to estimate γ. An alternative way to estimate the regression parameters γ and β is obtained by the quasi-likelihood. Also, according to Ma et al.(2009), with the Newton scoring algorithm, the value of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are:

$$\boldsymbol{g^*} = \boldsymbol{g} - S^{-1}(\boldsymbol{g})\psi(\boldsymbol{g}) \tag{3.55}$$

where

$$S = \frac{1}{4}\boldsymbol{X^T}\text{diag}E(\boldsymbol{Y})cov^{-1}(\boldsymbol{Y})\text{diag}E(\boldsymbol{Y})\boldsymbol{X^T} \tag{3.56}$$

$$g = [\gamma, \beta] \tag{3.57}$$

and

$$\psi(g) = [\psi(\gamma), \psi(\beta)] \tag{3.58}$$

Furthermore, a general matrix expression of $\psi(\boldsymbol{\gamma})$ and $\psi(\boldsymbol{\beta})$ can be expressed as:

$$\begin{aligned}\psi(g) &= \frac{1}{2}\boldsymbol{X^T}\text{diag}E(\boldsymbol{Y})cov^{-1}(\boldsymbol{Y})\{\boldsymbol{Y} - E(\boldsymbol{Y})\} \\ &= \frac{1}{2}\sum_{i=1}^{m}\boldsymbol{X_i^T}\text{diag}E(\boldsymbol{Y_i})cov^{-1}(\boldsymbol{Y_i})\{\boldsymbol{Y_i} - E(\boldsymbol{Y_i})\}\end{aligned} \tag{3.59}$$

3.4.2 Estimation of Random Effects Parameters

In the previous section, to estimate the two regression parameters γ and β, we considered the random effects parameters σ^2, τ^2 and ϵ^2 are known.

In this section, following "adjusted Pearson estimators", we are going to estimate the unknown random effects parameters σ^2, τ^2 and ϵ^2. According to Ma(2009), the Pearson estimators are adjusted by its bias correction. Hence, in general, the σ^2 can be expressed as:

$$\hat{\sigma}_r^2 = \frac{1}{m}\sum_{i=1}^{m}\{(\hat{U}_i - \frac{1}{2})^2 + c\} \tag{3.60}$$

where c is the bias correction. The bias correction c can be expressed as:

$$\begin{aligned} c &= E[\sigma^2 - \frac{1}{m}\sum_{i=1}^{m}(\hat{U}_i - \frac{1}{2})^2] \\ &= \sigma_{r-1}^2 - E[(\hat{U}_i - \frac{1}{2})^2] \\ &= \sigma_{r-1}^2 - \mathrm{Var}(\hat{U}_i) \\ &= \sigma_{r-1}^2 - \mathrm{Cov}(U_i, Y_i)\mathrm{Var}^{-1}(Y_i)\mathrm{Cov}(Y_i, U_i) \\ &= \sigma_{r-1}^2 - \sigma_{r-1}^4 \nu_i^{'}\mathrm{Var}^{-1}(Y_i)\nu_i \end{aligned} \tag{3.61}$$

With (3.55) and (3.56), the σ^2 can be simplified as:

$$\hat{\sigma}_r^2 = \frac{1}{m}\sum_{i=1}^{m}\{(\hat{U}_i - \frac{1}{2})^2 + \sigma_{r-1}^2 - \sigma_{r-1}^4 \nu_i^{'}\mathrm{Var}^{-1}(Y_i)\nu_i\} \tag{3.62}$$

where σ_{r-1}^2 is the estimate from the previous iteration. And by using the similar method, the τ^2 and ϵ^2 can be expressed as:

$$\hat{\tau}_r^2 = \frac{1}{T}\sum_{i=1}^{m}\sum_{j=1}^{N_i}\{(\hat{W}_{ij} - \hat{U}_i)^2 + \hat{\tau}_{r-1}^2 - [\sigma^4 \nu_i' \mathrm{Var}^{-1}(Y_i)\nu_i + \mathrm{Cov}(W_{ij}, Y_i)\mathrm{Var}^{-1}(Y_i)$$

$$\mathrm{Cov}(Y_i, W_{ij}) - 2\mathrm{Cov}(U_i, Y_i)\mathrm{Var}^{-1}(Y_i)\mathrm{Cov}(Y_i, W_{ij})\} \tag{3.63}$$

and

$$\hat{\epsilon}_r^2 = \frac{1}{T}\sum_{i=1}^{m}\sum_{j=1}^{N_i}\frac{2}{\mu_{ij}^2}\{(Y_{ij2} - \mu_{ij}\hat{W}_{ij})^2 + \mu_{ij}^2(\sigma^2 + \tau^2) - \mu_{ij}^2\mathrm{Cov}(W_{ij}, Y_i)\mathrm{Var}^{-1}(Y_i)\mathrm{Cov}(Y_i, W_{ij})\} \tag{3.64}$$

where T is total subject size and $\nu_i = (\pi_{i1}, ..., \pi_{iN_i}, \mu_{i1}, ..., \mu_{iN_i})'$

3.5 Parameters Initialization Computation Process

During the computation process, we need to initialize some parameters first. For example, the random effects U_i, the dispersion parameters σ^2 and even the regression parameters γ and β. Generally, different initial values will not make much difference to the final result. However, reasonable initial values can make the model more accurate and ultimately provide a better prediction result.

3.5.1 Initialize Regression Parameters

To initialize the regression parameter $\hat{\gamma}_{(0)}$, we first fit our data into generalize linear model with binomial distribution to get the fitted values. Let $\pi_{ij} =$ 2$*$fitted values. Then, using equation $log(\pi_{ij}/(1-\pi_{ij})) = logit(\pi_{ij})$ and another linear regression model $logit(\pi) = \gamma_0 + \gamma_1 * X_1 + \gamma_2 * X_2$ to get the regression parameters $\hat{\gamma}_{(0)}$. This step can be done by using glm function in R software or statsmodels package in Python software.

To initialize the regression parameter $\hat{\beta}_{(0)}$, we fit our data into generalize linear model with gamma distribution to get the coefficient and fitted values. The coefficient we got from the model will be our $\hat{\beta}_{(0)}$ except we will add another log(2) to the intercept term $\hat{\beta}_{0(0)}$.

3.5.2 Initialize Random Effects

Initial values for random effects $\hat{U}_{i(0)}$, $\hat{V}_{ij(0)}$ and $\hat{W}_{ij(0)}$ can be obtained by the following equations,

$$\hat{U}_{i(0)} = \begin{cases} \frac{1}{T}\sum_{i=1}^{m}\sum_{j=1}^{N_i} Y_{ij1}, & \frac{1}{N_i}\sum_{j=1}^{N_i} Y_{ij1} = 0|1 \\ \frac{1}{N_i}\sum_{j=1}^{N_i} Y_{ij1}, & \text{otherwise} \end{cases} \tag{3.65}$$

$$\hat{V}_{ij(0)} = \hat{U}_{ij(0)} \tag{3.66}$$

$$\hat{W}_{ij(0)} = \frac{Y_{ij2}}{\sum_{i=1}^{m}\sum_{j=1}^{N_i} Y_{ij2}/T} \tag{3.67}$$

The equation (3.65) shows that the initial value of U_i is equal to malformation proportion in the i^{th} litter. However if all subject's malformation is equal to 1 or 0 in the i^{th} litter. Then the initial value of U_i is equal to the malformation proportion of all observations.

The equation (3.66) shows that the initial value of V_i is equal to the initial value of U_i, if there is only one measurement on subject j within cluster i.

The equation (3.77) shows that the initial value of W_i is equal to the fetal weight on subject j within cluster i divided by the mean fetal weight of all baby mice.

3.5.3 Initialize Dispersion Parameters

Initial values for dispersion parameters $\hat{\sigma}^2$, $\hat{\tau}^2$ and $\hat{\epsilon}^2$ can be obtained by the following equations,

$$\hat{\sigma}^2_{(0)} = \frac{\sum_{i=1}^{m}(\hat{U}_i - E(\hat{U}_i))^2}{m} \tag{3.68}$$

$$\hat{\tau}^2_{(0)} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{N_i}(\hat{W}_{ij(0)} - \hat{U}_{ij(0)})^2}{T} \tag{3.69}$$

$$\hat{\epsilon}^2 = 1 \tag{3.70}$$

3.6 Update Iteration Process

Once all parameters are successfully initialized, then the update iteration process starts as follows,

1. Update the random effects $\hat{U}$, $\hat{V}$ and $\hat{W}$ by using the equation (3.30),

(3.35) and (3.43).

2. Update the regression parameters $\hat{\gamma}$ and $\hat{\beta}$ by using the equation (3.55).

3. Update the dispersion parameters $\hat{\sigma}^2$, $\hat{\tau}^2$ and $\hat{\epsilon}^2$ by using the equation (3.62), (3.63) and (3.64) respectively.

During the end of each iteration, calculate the absolute difference value $d_{(r)}$ between the current round r and previous round $r-1$, the absolute difference value $d_{(r)}$ can be expressed as:

$$d_{(r)} = |\hat{\beta}_r - \hat{\beta}_{r-1}| + |\hat{\gamma}_r - \hat{\gamma}_{r-1}| + |\hat{\sigma}_r^2 - \hat{\sigma}_{r-1}^2| + |\hat{\tau}_r^2 - \hat{\tau}_{r-1}^2| + |\hat{\epsilon}_r^2 - \hat{\epsilon}_{r-1}^2| \quad (3.71)$$

The update iteration process will stop if the absolute difference value is considerably small, here we use $d_{(r)} < 10^{-5}$.

Chapter 4

Data Analysis

The Tweedie mixed effects models are able to accommodate responses which follow Poisson distribution, gamma distribution, normal distribution or many more due to their flexibility. This chapter will focus on a model for the mice fetal dataset. The dataset has responses of two different types, the binary type response and the continuous type response. As previously discussed in Tweedie model, different values of the power index parameter indicate the different kinds of distribution. We estimated the index parameter based on the data, and the result is 2. Because gamma distribution corresponds index parameter equals to 2 in Tweedie. Therefore, we will use Bernoulli distribution for binary response and gamma distribution for continuous response. The detail of the dataset will be introduced in sections 4.1 and 4.2. Section 4.3 will focus on the model which we used to analyze the dataset. The results will be discussed in the next following section. Finally, the details of simulation process and simulation results will be included in the last

section.

4.1 Toxicity Study of Ethylene Glycol in Mice

The data is from a developmental toxicity study of ethylene glycol (EG). Ethylene glycol is a high-volume industrial chemical used in many applications. In a study of laboratory mice conducted through the National ToxicologyProgram (NTP), EG was administered at doses of 0, 750, 1500, or 3000 mg/kg/day to 94 pregnant mice (dams) over the period of major organogenesis, beginning just after implantation. The details of the original data can be found in table 4.1.

Table 4.1: The original data of the toxicity study of EG in mice

Variables	Description
Litter ID	ID number for each litter of baby mice, contains totally 94 litters.
Dose	The dose level of Ethylene glycol that applied to the mother mice. (0, 750, 1500, or 3000 mg/kg/day)
Fetal Weight	The baby mice fetal weight, a continuous response variable.
Fetal Malformations	The baby mice fetal malformation, a binary response variable. (0 stands for normal and 1 stands for malformation)

In this analysis, we study partial data containing only three dose levels (0, 750, 1500 mg/kg/day), having in total 802 observations on 71 litters. The

71 litters of baby mice are born from 71 pregnant mother mice. For convenience, the litter ID is coded as integers from 1 to 71. Each mother mice will give birth to a litter of mice. These 71 mother mice will use one of the three different doses. To avoid the scaling problem, the dose level was coded as 0 if the dose level is 0 mg/kg/day, 1 if the dose level is 750 mg/kg/day and 2 if the dose level is 1500mg/kg/day. Our attention will focus on those baby mice born from the mother mice who took the dose. The two responses were recorded. The first response is the fetal malformation which is a binary variable with 0 representing normal and 1 representing malformation. The second response is the fetal weight which is considered as a continuous variable reflecting the weight of the newly born baby mouse. In addition, each mother mouse will give birth to a different number of baby mice. The number of baby mice born from each mother mouse we assume follows a Poisson distribution with $\lambda = 11$. Figure 4.1 shows the percentage of each litter size with the Poisson probability mass function curve($\lambda = 11$). Table 4.2 shows the details of the data we used in this analysis.

4.2 Exploratory Data Analysis

Before attempting to specify a model for the data, it is best to perform some exploratory analysis using basic statistical methods. The first thing we can check is if the mean value of our response variables depends on the first

Figure 4.1: The percentage of each litter size with the Poisson probability mass function curve($\lambda = 11$)

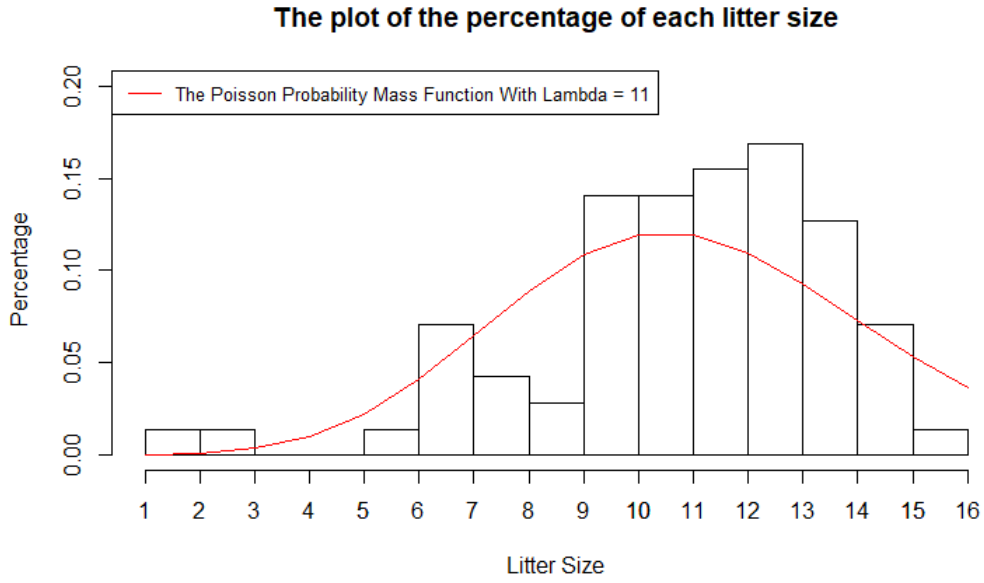


Table 4.2: A partial data of the toxicity study of EG in mice

Variables	Description
Litter ID	ID number for each litter of baby mice, contains totally 71 litters. (coded as integer from 1 to 71)
Dose	The dose level of Ethylene glycol that applied to the mother mice. (coded as 0, 1 and 2)
Fetal Weight	The baby mice fetal weight, a continuous response variable.
Fetal Malformations	The baby mice fetal malformation, a binary response variable. (0 stands for normal and 1 stands for malformation)
Litter Size	An integer indicating the size of the specific litter.

covariate "dose levels". Table 4.3 shows the average fetal weight and fetal malformation rate of baby mice at all different dose levels. This will give us an idea of how the dose levels affect the overall population.

Table 4.3: Fetal weight and fetal malformation rate based on three dose levels

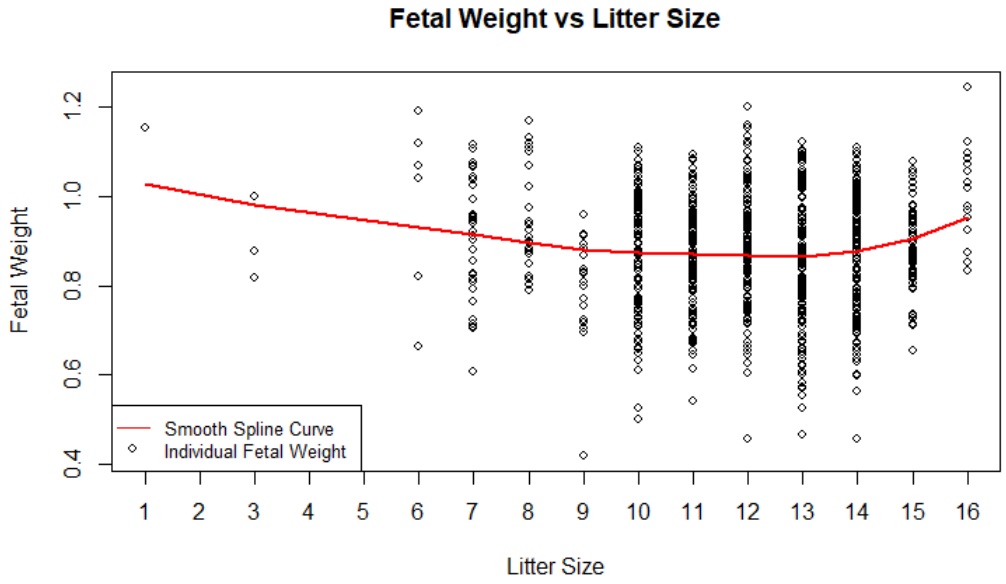
Dose level	Fetal Weight	Fetal Malformation Rate
Dose 0 (0mg/kg/day)	0.9719	0.0034
Dose 1 (750mg/kg/day)	0.8771	0.0942
Dose 2 (1500mg/kg/day)	0.7638	0.3886

From table 4.3, we can see that different dose levels will have different results on the fetal weight of baby mice and fetal malformation rate of baby mice. Also, from the results, it can be easily seen that the more ethylene glycol used in the mother mice, the less the fetal weight of their baby mice. Moreover, the more ethylene glycol used in the mother mice, the higher the fetal malformation rate of their baby mice. Overall, we can conclude that there is a potential negative correlation between dose levels and fetal weight. As well as a potential positive correlation between dose levels and fetal malformation.

Next, we can check if the value of our response variables depends on the second covariate, litter size. As we previously discussed in section 4.1, each mother mouse will give birth to a different number of baby mice. The number of baby mice born from each mother mouse we call it the litter

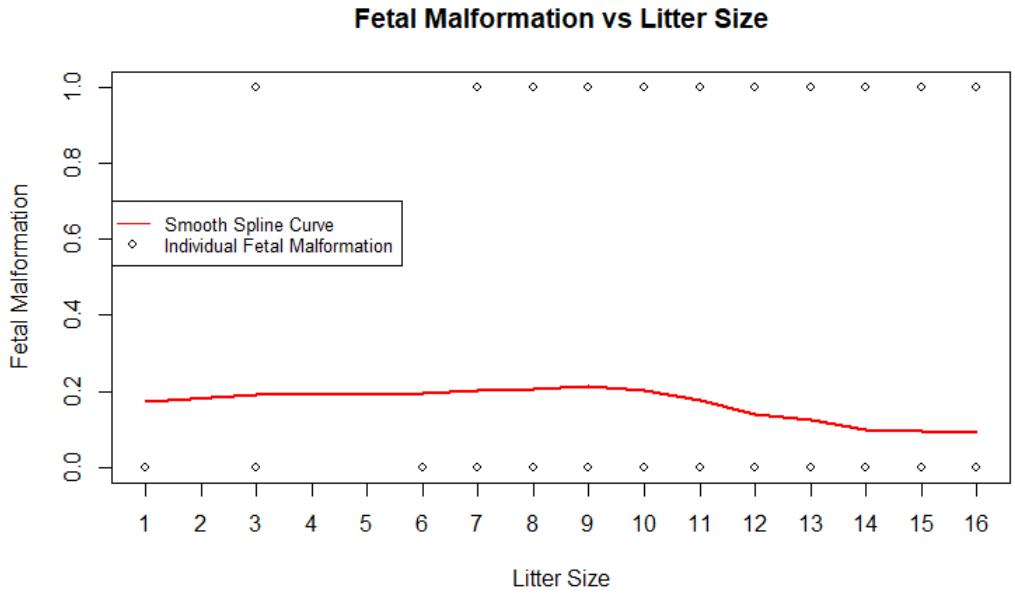
size. From the figure 4.2, the overall pattern among all points is not that obvious, however with a smooth spline curve, we can still see that as the litter size becomes larger, the fetal weight of the newborn mouse becomes smaller until the litter size 9 and then becomes larger to the end. So, there is a weak correlation between litter sizes and fetal weight.

Figure 4.2: Fetal weight based on Litter size with smooth spline curve



And from figure 4.3, the overall pattern seems the opposite of the previous fetal weight one. As the litter size becomes larger, the fetal malformation rate of the newborn mouse becomes higher until the litter size 9 and then becomes lower to the end. So, there is a weak correlation between litter sizes and fetal malformation as well.

Figure 4.3: Fetal malformation based on Litter size



4.3 Model Specification

For the binary response fetal malformation, we use Bernoulli distribution. Let Y_{ij1} be the j^{th} baby mouse malformation on i^{th} mother mouse. The first response Y_{ij1} is conditional on random effects $Z = (U, V, W)$. The random effects $U_1, ..., U_i, ..., U_m$ are the litter random effects and $V_{11}, ..., V_{ij}, ..., VmN_m$ are baby mouse subject-specific random effects for the baby mouse malformation. The conditional distribution of the response can be summarized as follows:

$$Y_{ij1}|Z \sim Ber(\pi_{ij}V_{ij}) \tag{4.1}$$

where $i = 1, ..., 71$ and $j = 1, ..., N_i$. In summary, the fetal malformation can be expressed as:

$$\text{logit}(\pi_{ij}) = x_{ij}^{'}\gamma \tag{4.2}$$

with the vector of covariates x_{ij}, measured at the binary response on the jth baby mouse subject within ith mother mouse. We consider dose level and the number of baby mice born form each mother mouse or in other words, litter size is covariates. As we previously discussed, in order to avoid the scaling problem, the covariate dose level was coded as 0 if the dose level is 0, 1 if the dose level is 750 and 2 if the dose level is 1500. The regression parameters γ_0, γ_1 and γ_2 are corresponding to intercept, dose level and litter size respectively.

For the continuous response fetal weight, we need to first select an optimal power index for Tweedie. We use the "tweedie.profile" function (Dunn, 2017) which built in R (R Core Team, 2019) to calculate the maximum likelihood estimate for a series of the power index values. The result of power index value is 2.0, so we use this value to fitted in Tweedie model. Also, this shows that our continuous response follows Gamma distribution. Let Y_{ij2} be the j^{th} baby mouse birth weight on i^{th} mother mouse. The second response Y_{ij2} is conditional on random effects $Z = (U, V, W)$. The random effects $U_1, ..., U_i, ..., U_m$ are the litter random effects and $W_{11}, ..., W_{ij}, ..., WmN_m$ are baby mouse subject-specific random effects for the baby mouse birth weight. The conditional distribution of the response can be summarized as

follows:

$$Y_{ij2}|Z \sim Tw_p(\mu_{ij}W_{ij}, \epsilon^2 W_{ij}^{1-p}) \tag{4.3}$$

where $i = 1, ..., 71$,$j = 1, ..., N_i$ and set Tweedie index parameter $p = 2$. In summary, the fetal weight can be expressed as:

$$\mu_{ij} = \exp(x_{ij}^{'}\beta) \tag{4.4}$$

with the vector of covariates x_{ij}, measured at the binary response on the jth baby mouse subject within ith mother mouse. Similar to the continuous response fetal weight, we still consider dose level and litter size are covariates. The regression parameters β_0, β_1 and β_2 are corresponding to intercept, dose level and litter size respectively.

4.4 Modelling Analysis Results

In this section, three methods will be used to analyze the toxicity study of EG in mice dataset. The generalized linear model, which is proposed by Nelder and Wedderburn (1972) will be discussed in 4.4.1. The univariate model, which is proposed by Ma et al. (2015) will be discussed in 4.4.2. The joint model which is covered by the previous section, the results will be discussed in 4.4.3.

4.4.1 Generalized Linear Model Analysis

The generalized linear model is a flexible generalization of ordinary linear regression, which allows error distribution models with the non-normal distribution of response variables. In our study, the binary response fetal malformation is assumed to follow Bernoulli distribution. In this situation, the ordinary linear regression is not suitable due to the response variable must be restricted to normal distribution. However, generalized linear models cover all these situations by allowing for response variables that have arbitrary distributions, which can be Poisson, Gamma and Binomial distribution. We use a generalized linear model to analyze the two responses separately.

Table 4.4: Parameter estimation results for Malformation (GLM)

	Fetal Malformations		
Parameters	Estimate	SE	P-value
Intercept	-2.6931	0.0545	<0.0001
Dose level	2.5514	0.0119	<0.0001
Litter size	-0.1032	0.0043	<0.0001

The estimate of regression parameters of the first response variable fetal malformations can be found in Table 4.4. All terms are significant due to their extremely small p-value. The estimate of dose level is 2.5514 which indicates the higher dose level may result in an increased chance of being malformation. The estimate of little size is negative value means that it is

negatively correlated with fetal malformation.

Table 4.5: Parameter estimation results for Weight (GLM)

	Fetal Weight		
Parameters	Estimate	SE	P-value
Intercept	0.7689	0.0238	<0.0001
Dose level	-0.1233	0.0052	<0.0001
Litter size	-0.0080	0.0019	<0.0001

The estimate of regression parameters of the second response variable fetal weight can be found in Table 4.5. Again, all terms are significant due to their extremely small p-value. The estimate of both dose level and little size are negative numbers, -0.1233 and -0.0080. Hence both terms are negatively correlated with fetal weight.

Although the GLM method can have some great results, it cannot analyze the two responses together nor can analyze the random effects.

4.4.2 Univariate Model Analysis

Similar to the GLM that we covered previously, the model here can handle the response variables that are not only normally distributed. Also, in order to analyze the random effects, the univariate model is using the generalized linear model with Tweedie distribution. The assumptions are the same as the joint model assumptions that we covered in section 3. But still, the

model needs to analyze the two responses separately.

Table 4.6: Parameter estimation results for Malformation (Univariate)

	Fetal Malformations		
Parameters	Estimate	SE	P-value
Intercept	-3.3188	1.4301	0.0203
Dose level	3.0353	0.5858	<0.0001
Litter size	-0.1057	0.1097	0.3353
σ^2	0.0847		

Table 4.6 shows the estimate of regression parameters and dispersion parameters of response fetal malformations. The p-value of dose level and little size is 0.3353, which indicates that the litter size is not significant. The estimate of the dose level is 3.0353, which is similar to the results of the generalized linear model. It is concluded that the higher dose level may result in an increased chance of being malformation.

Table 4.7: Parameter estimation results for Weight (Univariate)

	Fetal Weight		
Parameters	Estimate	SE	P-value
Intercept	0.8242	0.0508	<0.0001
Dose level	-0.1209	0.0130	<0.0001
Litter size	-0.0128	0.0040	0.0014
σ^2	0.0017		
τ^2	0.0009		
ϵ^2	0.0027		

Table 4.7 shows the estimate of regression parameters and dispersion parameters of response fetal weight. The overall results are similar to the results of the generalized linear model. The p-value of dose level and litter size is <0.0001 and 0.0014, therefore all terms are significant due to their small p-value. The estimate of both dose level and little size are negative numbers, -0.1209 and -0.0128. Hence both terms are negatively correlated with fetal weight.

4.4.3 Joint Model Analysis

Compare to the generalized linear model in section 4.4.1 and the univarite model in section 4.4.2, the joint model can not only analyze the random effects but also analyze the two response variables together. Therefore, the

covariance between the two responses will be taken into account.

By using the parameters updating iteration process which is covered in sections 3.5 and 3.6, the model successfully converged after 66 iterations. The results are shown in Table 4.8. From the results, the only term that not statistically significant is litter size for the response of fetal malformation, in which the p-value is 0.1471. All other terms are significant due to the very small p-value.

Table 4.8: Parameter estimation results (Joint Model)

	Fetal Malformation			Fetal Weight		
Parameters	Estimate	SE	P-value	Estimate	SE	P-value
Intercept	-2.9101	1.1410	0.0107	0.8224	0.0500	<0.0001
Dose level	2.9298	0.3866	<0.0001	-0.1206	0.0129	<0.0001
Litter size	-0.1325	0.0914	0.1471	-0.0126	0.0040	0.0016
σ^2	0.0017			0.0017		
τ^2				0.0009		
ϵ^2				0.0027		

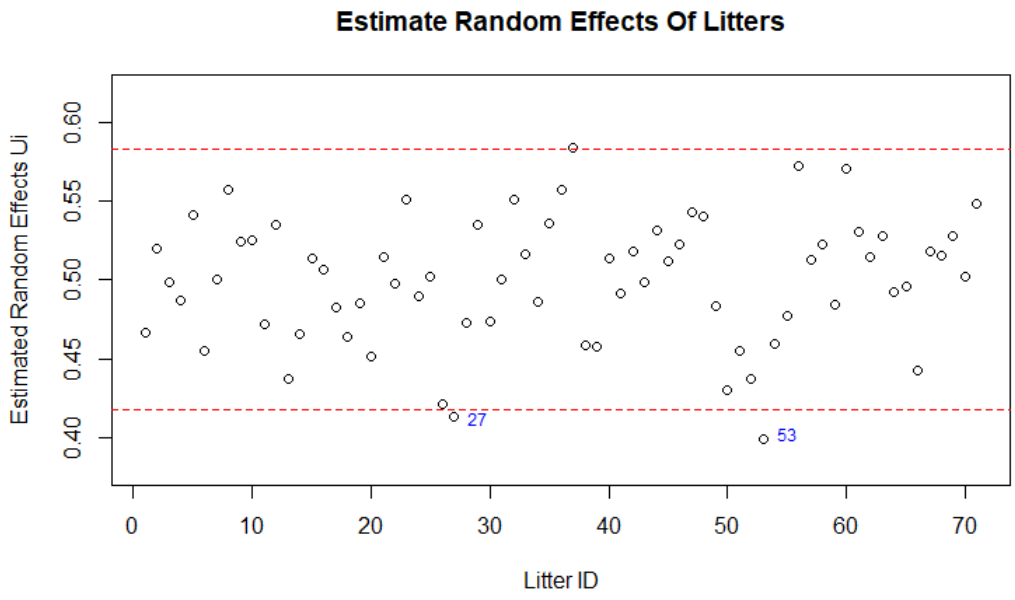
The estimate of the regression parameters γ_0, γ_1 and γ_2 for the fetal malformation response are -2.9101, 2.9298 and -0.1325 respectively. The estimate of the regression parameters β_0, β_1 and β_2 for the fetal weight response are 0.8224, -0.1206 and -0.0126 respectively.

The estimate of the dispersion parameters σ^2, τ^2 and ϵ^2 are 0.0017, 0.0009 and 0.0027 respectively.

It is also important to check if a specific litter deviates from all other litters.

In order to do that, the best way is to use the scatter plot of the estimated random effects of litters (U_i).

Figure 4.4: Cluster-specific random effects of each litter

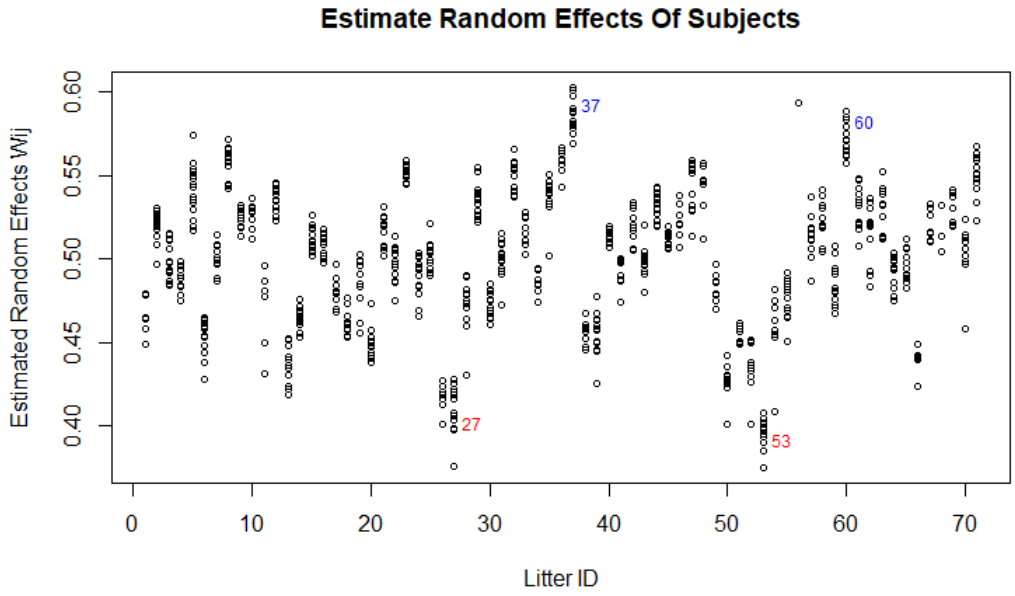


From figure 4.4, any points outside the red line are considered outliers. The red lines are calculated by 2 standard deviations of the mean $E(U_i)$. The plot indicates that there are two outliers, and litter 53 is considered to be an obvious outlier. It is clearly seen that the estimated random effect of litter 53 (U_{53}) is much lower than other points, which indicates that the baby mice from litter 53 are more likely to have smaller fetal weight. The same situation also applies to litter 27 but is not as obvious as litter 53.

Based on the model assumptions which is covered previously in chapter 3, the second level random effects V_{ij} are the same as the first level random

effects U_i if there is only one measurement on subject j within cluster i. So there is no need to check the subject-specific random effects on fetal malformation of each subject. However, the subject-specific random effects on fetal weight of each subject, which is the second random effects W_{ij} can be checked by using another diagram.

Figure 4.5: Subject-specific random effects Wij of each baby mouse



The estimated random effects on fetal weight of each baby mouse W_{ij} are shown on Figure 4.5. Each point represents the estimated random effects on fetal weight of each baby mouse, ordered by the litter i. It is clear to see that there is a huge variation among all the baby mice. But meanwhile, the estimated random effects are heavily dependent on the litter. For example, all estimate random effects on fetal weight of each baby mouse in litter 37

are larger than most of the other points. All estimate random effects on fetal weight of each baby mouse in litter 53 are smaller than most of the other points. Overall, despite the differences among each baby mouse, the baby mice from litter 37 and 60 are likely to have a larger fetal weight and the baby mice from litter 27 and 53 are likely to have a smaller fetal weight.

4.5 Simulation Study Results

In this section, simulation studies will be used for analyzing the overall model performance and evaluate the model bias and deviations. There are 1000 simulations in total. The simulated datasets are generated at the beginning of the simulation. In order to make the simulated data as close as possible to the real situation. The true parameters in the simulation study are those estimated parameters from the model and can be found in Table 4.8. The simulation datasets are generated by the following steps:

Step 1: Generate 100 litter sizes N_1,...,N_{100} following Poisson distribution with λ=11 for each litter.

Step 2: Generate 100 cluster-level random effects U_1,...,U_i,...,U_{100} following beta distribution with mean $\frac{1}{2}$ and variance σ^2.

Step 3: Generate N_i subject-level random effects V_{i1},...,V_{iN_i} for each V_i, following V_{i1},...,$V_{iN_i} = U_i$, where i = 1,..,100.

Step 4: Generate N_i subject-level random effects W_{i1},...,W_{iN_i} for

each W_i, following Gamma distribution with mean U_i and variance $\tau^2 U_i$, where i = 1,..,100.

Step 5: Generate simulated first level response Y_{111},...,Y_{ij1},...,$Y_{100N_{100}1}$ given two random effects U and V, following Bernoulli distribution with parameter p = $\pi_{ij}V_{ij}$, where $logit(\pi_{ij}) = x_{ij}^{'}\gamma$.

Step 6: Generate simulated second level response Y_{112},...,Y_{ij2},...,$Y_{100N_{100}2}$ given two random effects U and W, following Gamma distribution with mean $\mu_{ij}W_{ij}$ and variance $\epsilon^2\mu_{ij}^2 W_{ij}$, where $\mu_{ij} = \exp(x_{ij}^{'}\beta)$.

Step 7: Repeat step 1 to step 6 1000 times to generate 1000 simulated datasets.

Step 8: Once all simulated datasets are successfully generated, fits with the model we previous discussed and record all estimate results.

The overall simulated estimated value of the regression parameters, dispersion parameters as well as the simulation standard error can be calculated by taking the average of all 1000 simulated results. The final simulation summary results are shown in Table 4.9 and Table 4.10.

Table 4.9: Simulation Summary for the regression parameters in the toxicity study of EG in mice data

Parameters	True Value	Estimated Value	Bias	Estimated SE	Simulation SE
γ_0	-2.9101	-2.9781	-0.0680	0.6793	0.7035
γ_1	2.9298	3.0116	0.0818	0.3427	0.3612
γ_2	-0.1325	-0.1361	-0.0036	0.0550	0.0580
β_0	0.8224	0.8229	0.0005	0.0307	0.0311
β_1	-0.1206	-0.1204	0.0002	0.0101	0.0104
β_2	-0.0126	-0.0126	0.0000	0.0025	0.0026

Table 4.9 shows the simulation summary results of all regression parameters. The overall simulated estimate values of the regression parameters are listed in the Estimated Value column. The bias stands for the difference between the true value and the estimated value. From the results, the bias for all regression parameters are very small, as well as the standard errors.

Table 4.10 shows the simulation summary results of all dispersion parameters. The overall simulated estimate values of the dispersion parameters are listed in the Estimated Value column. From the results, the estimated value and the true value of the dispersion parameter σ^2 are very close. The dispersion parameters τ^2 and ϵ^2 are underestimated. Overall, according to the results of both Table 4.5 and Table 4.6, the previously estimated algorithm

Table 4.10: Simulation Summary for the dispersion parameters in the toxicity study of EG in mice data

Parameters	True Value	Estimated Value	Bias	Simulation SE
σ^2	0.0017	0.0016	-0.0001	0.0002
τ^2	0.0009	0.0003	-0.0006	<0.0001
ϵ^2	0.0027	0.0007	-0.0020	<0.0001

performs reasonably well.

Chapter 5

Discussion

5.1 Conclusion

In this study, the type of the two responses are completely different, one is the binary type and the other is the continuous type. Previously, there are many preexisting methods that could fit data with two responses. Generally, those methods are either analyze the two responses separately or do not consider random effects. In addition, preexisting methods are usually focused on one or more responses of the same type instead of different types. The proposed model in this thesis effectively fit data with two responses of different types by combining the Bernoulli and the Tweedie generalized model with shared random effects.

In this model, by using the two levels of random effects framework, the model is able to describe the cluster random effects between each cluster and the subject random effects between each subject within one cluster. The corre-

lation between each subject within one cluster is able to investigate as well. The random effects are predicted and estimated by using the Best Linear Unbiased Predictor (BLUP). There are some advantages to this model over the preexisting model. Compared to the generalized linear model, this proposed model can investigate both fixed effects and random effects instead of fixed effects only. Compared to the univariate model, this proposed model can handle the two different responses together instead of analyzing separately. Moreover, this model is very flexible as the Tweedie family of distributions are able to accommodate responses which follow many distributions such as Gamma distribution, Poisson distribution by using different power index values.

Finally, the overall model performance and the model bias and deviations are evaluated by the simulation. The bias for each parameter are very small and the estimated standard errors for each parameter are very close to simulation standard errors, which indicates the previously estimated algorithm and the model performs well.

5.2 Further Study

Although the results show that the model performs well, there are still many areas we could improve in the future.

In this thesis, we assume that the variance of second level random effects V_{ij} corresponding to the binary response given the first level random effects U_i is 0 if there is only one measurement at each cluster and each subject.

Therefore, unlike the second level random effects W_{ij}, there is no dispersion parameter for the second level random effects V_{ij}. The reason behind this is if we set a dispersion parameter τ_1 for random effects V_{ij}, then we will see that τ_1 does not exist in any cases of expression from 3.19 to 3.24 from the moment structure part. One possible solution is if we can combine two second level random effects V_{ij} and W_{ij} into one second level random effects, then the dispersion parameter τ will be shared by these two random effects. Next, the dose level of Ethylene glycol that is applied to the mother mice in the original data is 0, 750, 1500 or 3000 mg/kg/day. We have mentioned this previously in chapter 4, in order to avoid the scaling problem, the dose level was coded as 0 if the dose level is 0 mg/kg/day, 1 if the dose level is 750 mg/kg/day, 2 if the dose level is 1500mg/kg/day and 4 if the dose level is 3000mg/kg/day. But in fact, it is more reasonable to use the square root method ($\sqrt{\text{dose level}/750}$) instead of the linear method (dose level/750) to encode the dose levels. The reason is according to Fitzmaurice et al. (2010), because the decrease in fetal weight is not linear in increasing dose, but is approximately linear in increasing $\sqrt{dose}$. Also due to the starting point is 0 mg/kg/day instead of 1. Therefore, the dose level is more reasonable to be coded as 0 ($\sqrt{0}$) if the dose level is 0 mg/kg/day, 1 ($\sqrt{1}$) if the dose level is 750 mg/kg/day, 1.414 ($\sqrt{2}$) if the dose level is 1500mg/kg/day and 2 ($\sqrt{4}$) if the dose level is 3000mg/kg/day.

Although the model is very flexible, it still has some limitations. For example, the model only works when the binary proportion is less than half which is also the reason that we only use our model to focus on the partial data

without the dose level of 3000mg/kg/day. In the future, if the limitation of our model can be increased or the restriction can be completely removed, then the model will be much more flexible and may give out a more accurate results.

Bibliography

[1] Nelder, John; Wedderburn, Robert (1972). "Generalized Linear Models". *Journal of the Royal Statistical Society. Series A (General).* Blackwell Publishing. 135 (3): 370–384.

[2] Dunn, P.K., Smyth, G.K. (2005). Series evaluation of Tweedie exponential dispersion model densities. *Stat Comput* 15, 267–280.

[3] Peter McCullagh (1983). The Annals of Statistics, Mar., Vol. 11, No. 1 (Mar., 1983), pp. 59-67.

[4] D Hedeker (2005). "Generalized linear mixed models" - *Encyclopedia of statistics in behavioral science.*

[5] Breslow, N. E.; Clayton, D. G. (1993), "Approximate Inference in Generalized Linear Mixed Models".

[6] Robert D. Gibbons; D Hedeker; Stephen DuToit (2010). "Advances in Analysis of Longitudinal Data". *Annu Rev Clin Psychol.* 2010 Apr 27; 6: 79–107.

[7] Garrett M. Fitzmaurice; Nan M. Laird (1995) Regression Models for a Bivariate Discrete and Continuous Outcome with Clustering, *Journal of the American Statistical Association*, 90:431, 845-852.

[8] Lanjia Lin; Dipankar Bandyopadhyay; Stuart R. Lipsitz; Debajyoti Sinha (2010). "Association Models for Clustered Data with Binary and Continuous Responses". *Biometrics*, 66, 287-293.

[9] Garrett M. Fitzmaurice; NanM.Laird; James H. Ware (2010). *"Applied Longitudinal Analysis".* Chapter 22, 630-639.

[10] Jiming Jiang (2007). *"Linear and Generalized Linear Mixed Models and Their Applications".* Chapter 3, 204-205.

[11] Li, J (2018). Joint Tweedie Mixed Models for Longitudinal Data of Mixed Types. Master's thesis, University of New Brunswick.

[12] Jørgensen, B (1987). Exponential dispersion models. *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 49, no. 2, 127–162.

[13] Ma, R (1999). An Orthodox BLUP Approach to Generalized Linear Models. Ph.D. thesis, University of British Columbia.

[14] 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)*, vol. 69, no. 4, 625–641.

[15] Heather Turner (2008). Introduction to Generalized Linear Model *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 69, no. 4, 625–641.

[16] Ezzalfani, M; Burzykowski, T and Paoletti, X (2018). Joint modelling of a binary and a continuous outcome measured at two cycles to determine the optimal dose *Appl. Statist. (2019)*, vol. 68, no. 2, 369–384.

[17] Bekele, B.N.; Shen, Y. (2005). A Bayesian approach to jointly modeling toxicity and biomarker expression in a phase I/II dose-finding trial *Biometrics*, 61, 343-354.

[18] Lee, J.; Thall, P.F.; Y.; Mullerm P. (2015). Bayesian dose-finding in two treatment cycles based on the joint utility of efficacy and toxicity *J. Am. Statist. Ass*, 110, 711-722.

[19] Dunn, P. K. (2017). Tweedie: Evaluation of Tweedie exponential family models. *R package version 2.3.*

[20] R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. *URL https://www.R-project.org/*

Vita

Candidate's full name:

Tianyi Xia

Universities Attended:

Master of Science, August 2020, University of New Brunswick, Fredericton, NB, Canada

Bachelor of Science, December 2015, University of Toronto, Toronto, ON, Canada

Publications:

None

Conference Presentations:

None