2019

A MIXED-EFFECTS REGRESSION MODEL WITH APPLICATION TO LONGITUDINAL MISSING AT RANDOM MICROBIOME DATA

Manushi K.V. Welandawe
University of Rhode Island, manushividuneth@gmail.com

A MIXED-EFFECTS REGRESSION MODEL WITH APPLICATION TO LONGITUDINAL MISSING AT RANDOM MICROBIOME DATA

BY

MANUSHI K.V. WELANDAWE

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN

STATISTICS

UNIVERSITY OF RHODE ISLAND

2019

MASTER OF SCIENCE THESIS

OF

MANUSHI K.V. WELANDAWE

APPROVED:

Thesis Committee:

Major Professor Jing Wu

Gavino Puggioni

Yichi Zhang

Ashley Buchanan

Nasser H. Zawia

DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND

2019

ABSTRACT

Gut microbiota in the human lower gastrointestinal tract can impact the health through human functions as well as triggering numerous diseases. The colonization of the microbiota occurs immediately after the birth of an infant. Early life factors can highly influence the composition of gut microbiome, which ultimately affects infant's health. Therefore, the relationship between microbiome composition and clinical outcomes of preterm infants observed in Neonatal Intensive Care Units (NICUs) is of critical importance. To study this relationship, it is common to use longitudinal study designs. One of the major challenges in these designs is the huge percentage of missingness of the microbiome composition data, which needs to be appropriately accounted for during the study design. In this thesis, we propose a mixed-effects zero-inflated Beta regression model for longitudinal composition designs with missing at random data. This model captures the dependence of repeated measures for each subject by assuming a first-order autoregressive correlation structure. A Bayesian approach was employed for parameter estimations and inferences under this model. Performance of the model was investigated by a simulation study using different settings of missing data mechanisms. A sensitivity analysis was conducted to study the model misspecification issue. The developed model was further illustrated by a real data analysis on gut microbiome compositions of NICU preterm infants.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my major advisors Dr.Jing Wu and Dr.Gavino Puggioni for their encouragement, continuous support, patience, and motivation given to me throughout the entire study. Beside my advisors, I would like to thank my thesis committee members and chair: Dr.Yichi Zhang, Dr.Ashley Buchanan, and Dr.Aisling Caffrey for their insightful comments and encouragement. I am profoundly grateful to Dr.Amy D'Agata for permitting me to utilize the necessary data during this study. My heartfelt thanks go to Dr.Liliana Gonzalez, Dr.Natallia Katenka, and Dr.Prabhani Kuruppumullage for the guidance and support given to me. I would also like to thank my friends: TingFang Lee and Julie Osborn for their support and motivation provided throughout. A special thanks goes to my beloved husband for the unbroken encouragement, support, and motivation given to me throughout this study and throughout my life. I am immeasurably grateful to my parents and to my two sisters for their unstoppable encouragement, support, love, and care given to me throughout my life. Finally, I would like to thank all my colleagues, friends, and everyone who supported me in every step of this study to make it a success.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

Gut microbiota in the human lower gastrointestinal tract consists trillions of microorganisms and it accommodates approximately 150 times more genes than the entire human genomes [1, 2]. The bacterial cells in the gut is 10 times higher than the cells in the human body [3]. The gut mictobiota plays a major role in human health and disease [4]. Multiple studies reported the connection between the imbalance of gut microbiome and different human diseases such as obesity, inflammatory bowel disease, and diabetes [5, 6, 7, 8]. The exposure to psychosocial stressors could change the composition of these gut microbiota where this alteration can increase the vulnerability of an enteric pathogen [9, 10, 2].

Colonization of microbiota in the gastrointestinal tract begins immediately after the birth of an infant. Pattern of colonization for an infant will be affected based on the delivery method, the diet method, and environmental factors [11, 12]. This early life microbiome composition may affect the human health conditions in their later life stages [13, 14, 15]. This introduces the desirable need of understanding the microbiome composition of early life.

World Health Organization (WHO) defines a newborn as a preterm infant if the infant was born before the completion of thirty-seven weeks of the gestation period [16]. These preterm infants will be taken care at Neonatal Intensive Care Units (NICUs), which exposes them to different stressors. These stressors may result by the surrounding environment of the NICUs (eg: continuous bright lights and alarm/equipment noises) and painful caretaking procedures administered to the infants during their stay in NICUs [17, 18, 19]. Such exposure can affect the development of the infant's brain and could lead to learning difficulties in their later

life stages [20]. According to D'Agata et al. [21], this early life stress experienced in the NICU can influence the developing gut microbiome of preterm infants.

Most of the gut microbiome studies quantify the gut microbiota using DNA based methods such as high-throughput and low-cost sequencing methods [22]. A common approach is 16S ribosomal RNA (rRNA) gene sequencing method and the sequencing reads are used to identify the bacterial operational taxanomic units (OTUs or taxa) [23, 24]. The OTU read counts used to obtain the bacterial relative abundances that are bounded $[0, 1)$. A major challenge faced when conducting statistical analysis for microbiome composition data is the existence of excess zeros. This may caused due to the absence of the bacterial taxa in each sample [25].

Xia and Sun [26] reviewed different statistical methods which have been employed to analyze microbiome data. The two sample t-test and Wilcoxon rank sum test were used to test the species diversity in each individual sample between two groups [27, 28]. One-way analysis of variance (ANOVA) was applied to compare cutaneous microbiota of psoriatic group (affected group), unaffected group, and control group to perceive patterns that control skin colonization [29]. Voigt et al. [30] exerted a two-way ANOVA to observe taxonomic and functional specific biases initiated by RNALater preservation across all subjects. However, these standard statistical methods have their limitations on applying to longitudinal composition microbiome data [31]. For example, these methods does not account for the dependence of the repeated measures which occurs in longitudinal study designs.

Considering the excess amount of zeros present in OTU counts, Xu et al. [32] compared the performance of the zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB), and hurdle models by conducting extensive simulations with an application to a microbiome study. An additive logistic normal multinomial regression model was proposed by Xia et al. [33], which links covariates

to bacterial taxa counts with an application and tests the association between diet and stool microbiome composition. Li et al. [34] employed a multivariate zero-inflated logistic-normal model to analyze the association of disease risk factors with individual microbial taxa and overall microbial community composition. A zero-inflated generalized Dirichlet multinomial (ZIGDM) regression model was developed by Tang and Chen [35] to model the multivariate taxon counts.

To investigate how microbiome compositions change over time and how its trajectories relate to clinical outcomes, it is common to use longitudinal study designs [25, 36, 37]. In longitudinal study designs, the information is collected from the same subject repeatedly over a period of time. In these study designs, it is more often to confront intermittent missingness (non-monotone) and dropout (monotone) of subjects. The intermittent missingness may occur if a participant misses at least one visit during the study period and dropouts may occur if the participant withdraws from the study prematurely [38]. Little and Rubin [39] introduced three different types of missing data mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). The MCAR is where the missingness does not depend on the values of the observed or missing data, MAR is where the missingness depends only on observed data, and MNAR is where the missingness depends also on unobserved data. The MCAR and MAR are referred as ignorable missing data mechanism with the assumption that parameters of the missing data mechanism are distinct from the parameters of the sampling model. That is because the missing data mechanism does not need to be included in the likelihood function. On the other hand, MNAR is referred as non-ignorable missing data mechanism since the missing data mechanism cannot be omitted in the likelihood specification. A detailed description of the missing data mechanisms can be found in Chapter 2.

In statistical literature, several approaches have been proposed to handle different missing data mechanisms. The basic approach is casewise deletion where the cases with missing values in a variable(s) are deleted. The method is also known as listwise deletion or complete case analysis, which provides bias if the missing data mechanism is not MCAR [40]. Methods to incorporate MAR data mechanism include maximum likelihood approach using the expectation maximization (EM) algorithm, weighted generalized estimated equations, and multiple imputation [41, 42, 43]. Ibrahim et al. [44] reviewed four main approaches for handling missing covariate data in generalized linear models (GLM), i.e., maximum likelihood (ML), multiple imputation (MI), fully Bayesian (FB) and weighted estimating equations (WEE). Ibrahim et al. [38] developed a Monte Carlo EM (MCEM) algorithm to estimate parameters in generalized linear mixed models where the missing data mechanism is non-ignorable and missing data pattern is non-monotone.

Chen and Li [25] proposed a zero-inflated beta regression model (ZIBR) to study the relationship between the relative abundance of the microbiome with clinical covariates in a longitudinal study. This ZIBR model is a mixture of binary and beta regression components. The binary component models the presence of the bacterial genus and beta component models the non-zero microbiome composition. In their study, they included the random effects in the model to account for the correlation between repeated measurements of each subject in the study. However, their ZIBR model cannot handle the missing data encountered in longitudinal study designs. This ZIBR model was also used by D'Agata et al. [21] to assess the impact of early life stress on the trajectory of the gut microbial structure, where MCAR missing data mechanism was assumed. However, both ZIBR models proposed in the two studies assume that each individual shares the same random

effect among all repeated measures. This assumption could not always be satisfied because the random effect of an individual may not be the same in all the repeated measures of longitudinal study designs.

In this study, we developed a mixed-effects ZIBR model to understand the association between gut microbiome composition of the NICU preterm infants and their stress level when MAR missing data mechanism is assumed in the study design. This new model relaxes the assumption that each individual shares the same random effects in repeated measures by assuming first-order autoregressive covariance structure, AR(1). This correlation structure incorporates a high correlation between adjacent visits of an individual and the correlation systematically decreases when the distance between the visits increases. The Bayesian approach was used to estimate the parameters of this model using the Markov Chain Monte Carlo (MCMC) algorithm, which was implemented in *Nimble* statistical software [45]. Finally, the obtained results were compared with the study results of D'Agata et al. [21], which used the same data set for their analysis.

List of References

[1] I. Cho and M. J. Blaser, "The human microbiome: At the interface of health and disease," *Nature Reviews Genetics*, vol. 13, no. 4, p. 260, 2012.

[2] J. F. Cryan and T. G. Dinan, "Mind-altering microorganisms: The impact of the gut microbiota on brain and behaviour," *Nature Reviews Neuroscience*, vol. 13, no. 10, p. 701, 2012.

[3] F. Guarner and J.-R. Malagelada, "Gut flora in health and disease," *The Lancet*, vol. 361, no. 9356, pp. 512–519, 2003.

[4] M. J. Bull and N. T. Plummer, "Part 1: The human gut microbiome in health and disease," *Integrative Medicine: A Clinician's Journal*, vol. 13, no. 6, p. 17, 2014.

[5] P. J. Turnbaugh, R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis, and J. I. Gordon, "An obesity-associated gut microbiome with increased capacity for energy harvest," *Nature*, vol. 444, no. 7122, p. 1027, 2006.

[6] J. Qin, Y. Li, Z. Cai, S. Li, J. Zhu, F. Zhang, S. Liang, W. Zhang, Y. Guan, D. Shen, *et al.*, "A metagenome-wide association study of gut microbiota in type 2 diabetes," *Nature*, vol. 490, no. 7418, p. 55, 2012.

[7] C. Tamboli, C. Neut, P. Desreumaux, and J. Colombel, "Dysbiosis in inflammatory bowel disease," *Gut*, vol. 53, no. 1, pp. 1–4, 2004.

[8] B. Wang, M. Yao, L. Lv, Z. Ling, and L. Li, "The human microbiota in health and disease," *Engineering*, vol. 3, no. 1, pp. 71–82, 2017.

[9] R. T. Liu, "The microbiome as a novel paradigm in studying stress and mental health," *The American Psychologist*, vol. 72, no. 7, pp. 655–667, Oct 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29016169

[10] M. T. Bailey, S. E. Dowd, N. M. Parry, J. D. Galley, D. B. Schauer, and M. Lyte, "Stressor exposure disrupts commensal microbial populations in the intestines and leads to increased colonization by citrobacter rodentium," *Infection and Immunity*, vol. 78, no. 4, pp. 1509–1519, 2010.

[11] M. Carabotti, A. Scirocco, M. A. Maselli, and C. Severi, "The gut-brain axis: Interactions between enteric microbiota, central, and enteric nervous systems," *Annals of Gastroenterology: Quarterly Publication of the Hellenic Society of Gastroenterology*, vol. 28, no. 2, p. 203, 2015.

[12] J. A. Foster and K.-A. M. Neufeld, "Gut-brain axis: How the microbiome influences anxiety and depression," *Trends in Neurosciences*, vol. 36, no. 5, pp. 305 – 312, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0166223613000088

[13] I. Cho, S. Yamanishi, L. Cox, B. A. Methé, J. Zavadil, K. Li, Z. Gao, D. Mahana, K. Raju, I. Teitler, *et al.*, "Antibiotics in early life alter the murine colonic microbiome and adiposity," *Nature*, vol. 488, no. 7413, p. 621, 2012.

[14] R. A. Dimmitt, E. M. Staley, G. Chuang, S. C. Tanner, T. D. Soltau, and R. G. Lorenz, "The role of postnatal acquisition of the intestinal microbiome in the early development of immune function," *Journal of Pediatric Gastroenterology and Nutrition*, vol. 51, no. 3, p. 262, 2010.

[15] M. J. Ege, M. Mayer, A.-C. Normand, J. Genuneit, W. O. Cookson, C. Braun-Fahrländer, D. Heederik, R. Piarroux, and E. von Mutius, "Exposure to environmental microorganisms and childhood asthma," *New England Journal of Medicine*, vol. 364, no. 8, pp. 701–709, 2011.

[16] WHO, "Preterm birth," http://www.who.int/news-room/fact-sheets/detail/preterm-birth, 2018. [Online]. Available: http://www.who.int/news-room/fact-sheets/detail/preterm-birth

[17] N. Witt, S. Coynor, C. Edwards, and H. Bradshaw, "A guide to pain assessment and management in the neonate," *Current Emergency and Hospital Medicine Reports*, vol. 4, no. 1, pp. 1–10, 2016.

[18] T. Field, "Alleviating stress in newborn infants in the intensive care unit," *Clinics in Perinatology*, vol. 17, no. 1, pp. 1–9, 1990.

[19] R. E. Grunau, L. Holsti, D. W. Haley, T. Oberlander, J. Weinberg, A. Solimano, M. F. Whitfield, C. Fitzgerald, and W. Yu, "Neonatal procedural pain exposure predicts lower cortisol and behavioral reactivity in preterm infants in the NICU," *Pain*, vol. 113, no. 3, pp. 293–300, 2005.

[20] L. J. Woodward, J. O. Edgin, D. Thompson, and T. E. Inder, "Object working memory deficits predicted by early brain injury and development in the preterm infant," *Brain*, vol. 128, no. 11, pp. 2578–2587, 2005.

[21] A. L. DAgata, J. Wu, M. K. Welandawe, S. V. Dutra, B. Kane, and M. W. Groer, "Effects of early life NICU stress on the developing gut microbiome," *Developmental Psychobiology*, 2019.

[22] E. Thursby and N. Juge, "Introduction to the human gut microbiota," *Biochemical Journal*, vol. 474, no. 11, pp. 1823–1836, 2017.

[23] P. J. McMurdie and S. Holmes, "phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data," *PloS One*, vol. 8, no. 4, p. e61217, 2013.

[24] A. Dhariwal, J. Chong, S. Habib, I. L. King, L. B. Agellon, and J. Xia, "MicrobiomeAnalyst: A web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data," *Nucleic Acids Research*, vol. 45, no. W1, pp. W180–W188, 2017.

[25] E. Z. Chen and H. Li, "A two-part mixed-effects model for analyzing longitudinal microbiome compositional data," *Bioinformatics*, vol. 32, no. 17, pp. 2611–2617, 2016.

[26] Y. Xia and J. Sun, "Hypothesis testing and statistical analysis of microbiome," *Genes & Diseases*, vol. 4, no. 3, pp. 138–148, 2017.

[27] P. S. La Rosa, Y. Zhou, E. Sodergren, G. Weinstock, and W. D. Shannon, "Hypothesis testing of metagenomic data," in *Metagenomics for Microbiology*. Elsevier, 2015, pp. 81–96.

[28] X. Yin, J. Peng, L. Zhao, Y. Yu, X. Zhang, P. Liu, Q. Feng, Y. Hu, and X. Pang, "Structural changes of gut microbiota in a rat non-alcoholic fatty liver disease model treated with a chinese herbal formula," *Systematic and Applied Microbiology*, vol. 36, no. 3, pp. 188–196, 2013.

[29] A. V. Alekseyenko, G. I. Perez-Perez, A. De Souza, B. Strober, Z. Gao, M. Bihan, K. Li, B. A. Methé, and M. J. Blaser, "Community differentiation of the cutaneous microbiota in psoriasis," *Microbiome*, vol. 1, no. 1, p. 31, 2013.

[30] A. Y. Voigt, P. I. Costea, J. R. Kultima, S. S. Li, G. Zeller, S. Sunagawa, and P. Bork, "Temporal and technical variability of human gut metagenomes," *Genome Biology*, vol. 16, no. 1, p. 73, 2015.

[31] S. Mandal, W. Van Treuren, R. A. White, M. Eggesbø, R. Knight, and S. D. Peddada, "Analysis of composition of microbiomes: A novel method for studying microbial composition," *Microbial Ecology in Health and Disease*, vol. 26, no. 1, p. 27663, 2015.

[32] L. Xu, A. D. Paterson, W. Turpin, and W. Xu, "Assessment and selection of competing models for zero-inflated microbiome data," *PloS One*, vol. 10, no. 7, p. e0129606, 2015.

[33] F. Xia, J. Chen, W. K. Fung, and H. Li, "A logistic normal multinomial regression model for microbiome compositional data analysis," *Biometrics*, vol. 69, no. 4, pp. 1053–1063, 2013.

[34] Z. Li, K. Lee, M. R. Karagas, J. C. Madan, A. G. Hoen, A. J. OMalley, and H. Li, "Conditional regression based on a multivariate zero-inflated logistic-normal model for microbiome relative abundance data," *Statistics in Biosciences*, pp. 1–22, 2018.

[35] Z.-Z. Tang and G. Chen, "Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis," *Biostatistics*, 2018.

[36] J. D. Lewis, E. Z. Chen, R. N. Baldassano, A. R. Otley, A. M. Griffiths, D. Lee, K. Bittinger, A. Bailey, E. S. Friedman, C. Hoffmann, *et al.*, "Inflammation, antibiotics, and diet as environmental stressors of the gut microbiome in pediatric crohns disease," *Cell Host & Microbe*, vol. 18, no. 4, pp. 489–500, 2015.

[37] A. Gonzalez, A. King, M. S. Robeson II, S. Song, A. Shade, J. L. Metcalf, and R. Knight, "Characterizing microbial communities through space and time," *Current Opinion in Biotechnology*, vol. 23, no. 3, pp. 431–436, 2012.

[38] J. G. Ibrahim, M.-H. Chen, and S. R. Lipsitz, "Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable," *Biometrika*, vol. 88, no. 2, pp. 551–564, 2001.

[39] R. J. Little and D. Rubin, "Statistical analysis with missing data," *New York*, 2002.

[40] T. A. Myers, "Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data," *Communication Methods and Measures*, vol. 5, no. 4, pp. 297–310, 2011.

[41] D. B. Rubin, *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, 2004, vol. 81.

[42] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

[43] J. M. Robins, A. Rotnitzky, and L. P. Zhao, "Analysis of semiparametric regression models for repeated outcomes in the presence of missing data," *Journal of the American Statistical Association*, vol. 90, no. 429, pp. 106–121, 1995.

[44] J. G. Ibrahim, M.-H. Chen, S. R. Lipsitz, and A. H. Herring, "Missing-data methods for generalized linear models: A comparative review," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 332–346, 2005.

[45] NIMBLE Development Team, "NIMBLE: MCMC, particle filtering, and programmable hierarchical modeling," 2019. [Online]. Available: https://cran.r-project.org/package=nimble

CHAPTER 2

Methodology

2.1 Motivating Dataset

The motivating data set was obtained from D'Agata et al. study [1]. The data were originally collected from very low birth weight (VLBW) infants who were admitted to the NICU of Tampa General Hospital, Florida between the years 2011 and 2015. In the data collection process, they excluded infants whose birth weights were greater than 1500g, infants who had major congenital anomalies, moribund infants, and infants whose mothers were HIV-infected. A total of 82 VLBW infants were included in the final study sample after the exclusion procedure. A detailed description of the data collection method can be found in D'Agata et al. [1].

According to D'Agata et al. [1], the stools samples of VLBW infants were collected for six consecutive weeks since they were admitted to the NICU. After conducting the microbiome analysis, bacterial operational taxonomic units (OTUs) were reported at the genus level. These reported OTUs were used to calculate the bacterial relative abundance (microbiome composition), which was the dependent variable of this analysis. The infants' stress experience was recorded daily for 6 weeks since the admission to the NICU. These data were collected using the Neonatal Infant Stressor Scale (NISS), a quantitative instrument which measures the interventions performed in NICU [2]. The time dependent variables, stress scores two weeks (lag 2) and one week (lag 1) prior to the current week sampling and the sampling week were considered in the analysis as independent variables. Other covariates considered includes the baseline variables, infant's gender, birth weight, gestational age, and antibiotic usage. The weekly NISS stress scores of each infant are shown in Figure 1. The average weekly stress scores (shown by the black points) decrease over time, suggesting that infants have higher stress during

the first couple of weeks in the NICU.

In some instances, during this longitudinal study, some infants had multiple samples while some infants had no samples collected during certain weeks. We thus used the weekly average of the microbiome composition of the samples as the dependent variables. However, missing data were encountered when no samples were collected during the entire week. Due to the procedure of the data collection [1], MAR missing data mechanism was assumed throughout the study. After excluding the infants with missing covariates, the study sample was reduced to 68. Figure 2 shows the weekly missing data pattern of the microbial compositions for the samples of 68 infants. There is a higher number of missing values in the first week (67.7%), second week (30.9%), and sixth week (30.9%) compared to other three weeks in the study sample. We consider data from week 3 to week 6 during the analysis due to the higher occurrence of missingness in the first two weeks, as in D'Agata et al. [1] The overall missing data percentage of the response variable was 22.05% and all the covariates of the 68 infants were completely observed. Finally, after removing the low abundant genera from the obtained sample [1], 21 bacterial genera were included in the analysis.

2.2 Model for Longitudinal Microbiome Data

The microbiome composition of each bacterial genus is bounded from [0,1), zero-inflated, and the distribution is highly skewed (see Figure 3) [3]. Therefore, the ZIBR model with random effects in the beta component was used to study the association of gut microbiome composition and NICU stress of preterm infants.

Let y_t be the relative abundance for a given bacterial genus at time t, for $t = 0, 1, \ldots, T$, where y_0 represents the baseline measurement. Conditioning on the subject-level random effects η_t, we assume that y_t $(0 \leq y_t < 1)$ follows zero-

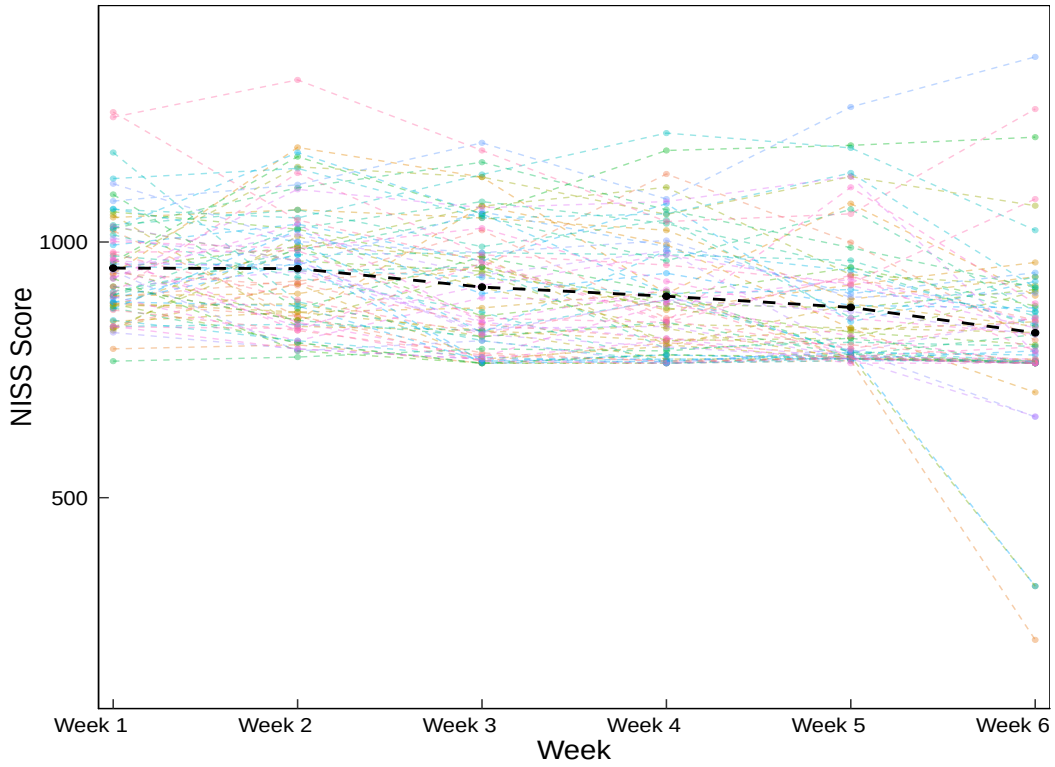


Figure 1: The black points shows the average score of each week and all other points represent the stress score of each infant.

inflated Beta distribution given by

$$f(y_t|\mathbf{x}_t, \mathbf{z}_t, \eta_t) = \begin{cases} 1 - p_t, & \text{if } y_t = 0 \\ p_t \cdot \frac{\Gamma(\phi)}{\Gamma(\mu_t\phi)\Gamma((1-\mu_t)\phi)} y_t^{\mu_t\phi-1}(1-y_t)^{(1-\mu_t)\phi-1}, & \text{if } y_t > 0, \end{cases} \tag{1}$$

and

$$\begin{aligned} \log(\frac{p_t}{1-p_t}) &= \mathbf{x}_t^\top \boldsymbol{\beta}, \\ \log(\frac{\mu_t}{1-\mu_t}) &= \mathbf{z}_t^\top \boldsymbol{\gamma} + \eta_t, \end{aligned} \tag{2}$$

where $(1 - p_t)$ is the probability of extra zeros, μ_t $(0 < \mu_t < 1)$ and ϕ $(\phi > 0)$ are the mean and the precision parameters of the Beta distribution, respectively. Let $\mathbf{x}_t$ and $\mathbf{z}_t$ are the two vectors of baseline covariates at time t. The $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_q)^T$ are the vectors of regression coefficients

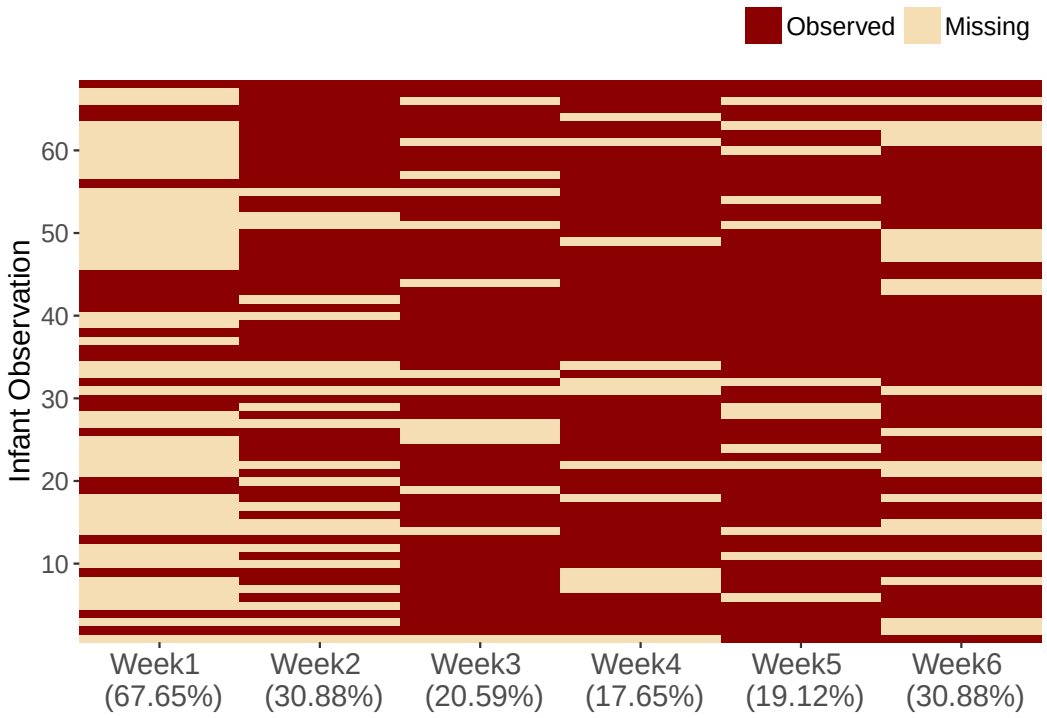


Figure 2: The missing and observed data represented by beige and maroon colors respectively.

corresponding to $\mathbf{x}_t$ and $\mathbf{z}_t$ where p and q are the dimensions of each vector. Let $\boldsymbol{\eta} = (\eta_0, \ldots, \eta_T)^T \sim N(\mathbf{0}, \sigma^2\Sigma)$, where $\boldsymbol{\Sigma}$ is a $(T+1) \times (T+1)$ correlation matrix with the (a, b) entries of the matrix being $\rho^{|b-a|}$ $(-1 < \rho < 1)$.

2.3 Missing Data Mechanism

Incomplete data are often encountered in many fields of studies. The missingness might be caused by various reasons, such as non-responses in survey questions, study drop-outs or failure to follow-up, instrumentation drawbacks and so forth. The three missing data mechanisms reviewed by Little and Rubin [4] which is mentioned in Chapter 1 can be shown using mathematical notations as follows.

Let Y, R, Y_{obs}, and Y_{mis} denote the complete data, missing data indicator matrix, observed, and missing measures of Y, respectively. The missing data

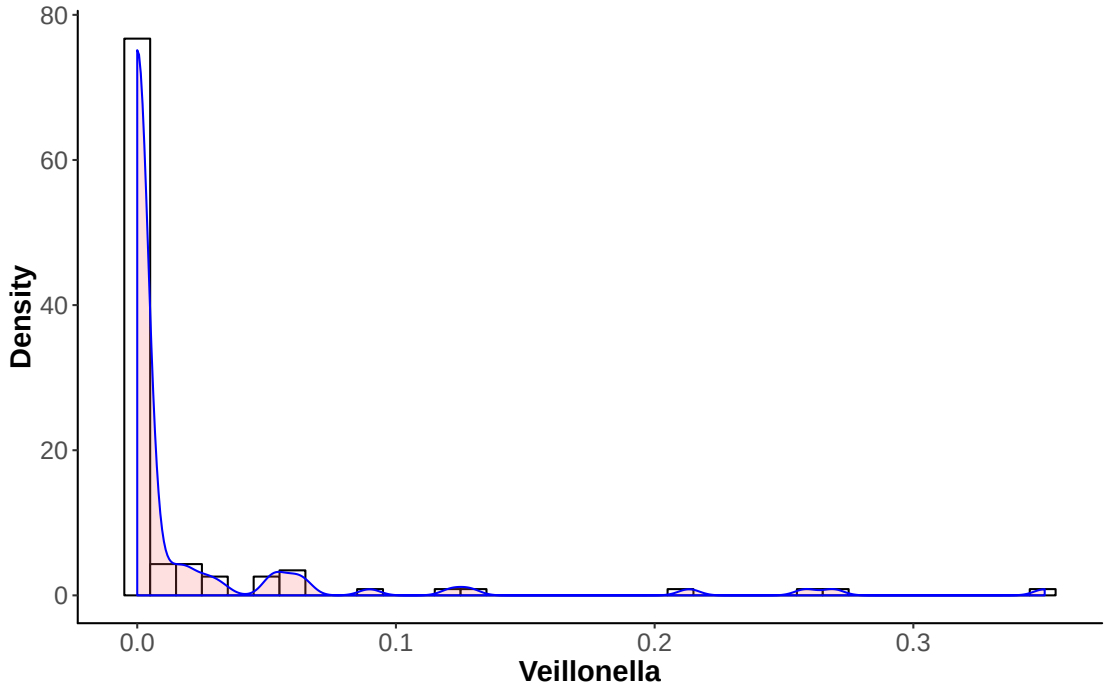


Figure 3: Bars represent the histogram and the curve represent the density estimation.

mechanism is given by $f(R \mid Y, \phi)$, the conditional distribution of R given Y and ϕ, where ϕ is the unknown parameter vector. Then the MCAR mechanism can be denoted as,

$$f(R \mid Y, \phi) = f(R \mid \phi) \text{ for all } Y, \phi, \tag{3}$$

where the missingness does not depend on the values of the observed or missing Y. The MAR mechanism can be denoted as,

$$f(R \mid Y, \phi) = f(R \mid Y_{obs}, \phi) \text{ for all } Y_{mis}, \phi, \tag{4}$$

where missingness depends only on observed data. Finally, the MNAR mechanism is when the missingness (R) depends also on the missing values of Y (Y_{mis}).

Let the joint probability distribution of Y and R given by

$$f(Y_{obs}, Y_{mis}, R \mid \theta, \phi) = f(Y_{obs}, Y_{mis} \mid \theta) f(R \mid Y_{obs}, Y_{mis}, \phi), \tag{5}$$

where θ is the vector of parameters of interest.

Then the joint probability distribution of observed data (Y_{obs}) and R can be obtained by integrating out Y_{mis}.

$$f(Y_{obs}, R \mid \theta, \phi) = \int f(Y_{obs}, Y_{mis} \mid \theta) f(R \mid Y_{obs}, Y_{mis}, \phi) dY_{mis}. \tag{6}$$

Let N denote the total number of subjects then the full likelihood function of θ and ϕ is given by

$$L_{full}(\theta, \phi \mid Y_{obs}, R) = \prod_{i=1}^{N} f(Y_{i,obs}, R \mid \theta, \phi) \quad (\theta, \phi) \; \epsilon \; \Omega_{\theta,\phi}, \tag{7}$$

where $i = (1, \ldots, N)$ denotes the i^{th} subject and Ω is the parameter space.

The likelihood function of θ ignoring the missing data mechanism is given by

$$L_{ign}(\theta \mid Y_{obs}) = \prod_{i=1}^{N} f(Y_{i,obs} \mid \theta). \tag{8}$$

If the distribution of the missing data mechanism does not depend on the unobserved Y_{mis} i.e., MAR (see (4)) and θ and ϕ are priori independent,

$$\pi(\theta, \phi) = \pi(\theta)\pi(\phi),$$

then (6) can be simplified as follows,

$$f(Y_{obs}, R \mid \theta, \phi) = f(Y_{obs} \mid \theta) f(R \mid Y_{obs}, \phi). \tag{9}$$

Finally, the likelihood-based inference for Y from $L_{full}(\theta, \phi \mid Y_{obs}, R)$ will be the same as the likelihood-based inference for Y from $L_{ign}(\theta \mid Y_{obs})$.

2.4 Bayesian Inference

Unlike the frequentist framework, where the parameters are considered as unknown constants, the parameters in the Bayesian framework are considered as random variables. As given by Gelman et al. [5], let θ be the parameter vector of interest and y be the vector of data obtained. By using the Baye's rule, the posterior density can be given by

$$\pi(\theta \mid y) = \frac{\pi(\theta, y)}{\pi(y)} = \frac{\pi(y \mid \theta)\pi(\theta)}{\pi(y)}, \tag{10}$$

where $\pi(\theta)$, $\pi(y \mid \theta)$, and $\pi(y) = \int \pi(\theta)\pi(y \mid \theta)d\theta$ are the prior distribution, sampling distribution, and the normalizing constant respectively. Since $\pi(y)$ does not depend on θ, (10) can be written as,

$$\pi(\theta \mid y) \propto \pi(y \mid \theta)\pi(\theta) \tag{11}$$

Therefore, the posterior density can be obtained by the likelihood function of parameters times the prior distribution of the parameters.

2.4.1 The Likelihood Function

Write $\mathbf{y}_i = (y_{i0}, \ldots, y_{iT})^\top$, $\mathbf{x}_i = (x_{i0}, \ldots, x_{iT})^\top$, and $\boldsymbol{\eta}_i = (\eta_{i0}, \ldots, \eta_{iT})^\top$. Let $r_i = (r_{i0}, \ldots, r_{iT})$ be the vector of missing data indicators, where r_{it} is given by

$$r_{it} = \begin{cases} 0, & \text{if } y_t \text{ is observed} \\ 1, & \text{if } y_t \text{ is missing} \end{cases} \tag{12}$$

Let $\mathbf{y}_{i,\text{obs}}$ and $\mathbf{y}_{i,\text{mis}}$ denote the observed and the missing measures for the i^{th} subject where $\mathbf{y}_{\text{obs}} = (\mathbf{y}_{1,\text{obs}}, \ldots, \mathbf{y}_{N,\text{obs}})^\top$ and $\mathbf{y}_{\text{mis}} = (\mathbf{y}_{1,\text{mis}}, \ldots, \mathbf{y}_{N,\text{mis}})^\top$.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \phi, \sigma, \rho)$ be the vector of all parameters. Then the likelihood function is given by

$$\begin{aligned} L(\boldsymbol{\theta} \mid \mathbf{y}) &= \int \prod_{i=1}^{N} f(\mathbf{y}_i|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\eta}_i, \boldsymbol{\theta}) f(\boldsymbol{\eta}_i|\sigma^2, \rho) d\boldsymbol{\eta} \\ &= \int \prod_{i=1}^{N} \Bigg[\prod_{t=0}^{T} (1 - p_{it})^{\mathbf{I}(y_{it}=0)} p_{it}^{\mathbf{I}(y_{it}>0)} \\ &\quad \left\{ \frac{\Gamma(\phi)}{\Gamma(\mu_{it}\phi)\Gamma((1-\mu_{it})\phi)} y_{it}^{\mu_{it}\phi-1} (1 - y_{it})^{(1-\mu_{it})\phi-1} \right\}^{\mathbf{I}(y_{it}>0)} \Bigg] \\ &\quad \left[\frac{1}{(2\pi\sigma^2)^{n_i/2} \mid \Sigma \mid^{1/2}} exp\left(\frac{-1}{2\sigma^2} \boldsymbol{\eta}_i^T \Sigma^{-1} \boldsymbol{\eta}_i \right) \right] d\boldsymbol{\eta}, \end{aligned} \tag{13}$$

where $n_i = (1, 2, \ldots, T+1)$ denote the number of observed values of each subject i.

By the assumption of missing at random data, the likelihood function will be given by

$$
\begin{aligned}
L(\boldsymbol{\theta} \mid \mathbf{y}_{obs}) &= \int \prod_{i=1}^{N} f(\mathbf{y}_i|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\eta}_i, \boldsymbol{\theta}) f(\boldsymbol{\eta}_i|\sigma^2, \rho) dy_{i,miss} d\boldsymbol{\eta} \\
&= \int \prod_{i=1}^{N} \prod_{t=0}^{T} \Bigg[(1-p_{it})^{\mathbf{I}(y_{it}=0)} (p_{it})^{\mathbf{I}(y_{it}>0)} \\
&\quad \left\{ \frac{\Gamma(\phi)}{\Gamma(\mu_{it}\phi)\Gamma((1-\mu_{it})\phi)} y_{it}^{\mu_{it}\phi-1} (1-y_{it})^{(1-\mu_{it})\phi-1} \right\}^{\mathbf{I}(y_{it}>0)} \Bigg]^{\mathbf{I}(r_{it}=0)} \times \\
&\quad \left[\frac{1}{(2\pi\sigma^2)^{n_i/2}} \frac{1}{\mid \mathbf{P}_i \Sigma \mathbf{P}_i^T \mid^{1/2}} exp\left(\frac{-1}{2\sigma^2} (\mathbf{P}_i \boldsymbol{\eta}_i)^T (\mathbf{P}_i \Sigma \mathbf{P}_i^T)^{-1} (\mathbf{P}_i \boldsymbol{\eta}_i) \right) \right] d\boldsymbol{\eta},
\end{aligned}
\tag{14}
$$

where $\mathbf{P}_i$ denote the $n_i \times (T+1)$ selection matrix to select the observed values.

2.4.2 Prior Distributions

The parameters of interest in this model are $\boldsymbol{\beta}, \boldsymbol{\gamma}, \phi, \sigma^2,$ and ρ. In the Bayesian analysis, we used non-informative priors for parameters. The proposed prior distributions for the coefficients of fixed effects of binary and beta components are multivariate normal: $\boldsymbol{\beta} \sim N_p(\mathbf{0}, 10^4 \mathbf{I}_p)$ and $\boldsymbol{\gamma} \sim N_q(\mathbf{0}, 10^4 \mathbf{I}_q)$ where $\mathbf{I}_p$ and $\mathbf{I}_q$ are $p \times p$ and $q \times q$ identity matrices respectively. For the precision parameter, we proposed an inverse gamma distribution: $\phi \sim IG(10^{-3}, 10^{-3})$. The variance and correlation parameters of the random effects, we used inverse gamma distribution: $\sigma^2 \sim IG(10^{-3}, 10^{-3})$ and uniform distribution: $\rho \sim U(0, 1)$.

The Bayesian analysis was conducted using the proposed prior distributions in *R* version 3.5.3 [6] using the *nimble* package version 0.8.0 [7, 8, 9]. This software uses Markov Chain Monte Carlo (MCMC) algorithms to obtain the samples from posterior distributions of the parameters.

List of References

[1] A. L. DAgata, J. Wu, M. K. Welandawe, S. V. Dutra, B. Kane, and M. W. Groer, "Effects of early life NICU stress on the developing gut microbiome," *Developmental Psychobiology*, 2019.

[2] C. A. Newnham, T. Inder, and J. Milgrom, "Measuring preterm cumulative stressors within the NICU: The neonatal infant stressor scale," *Early Human Development*, vol. 85, no. 9, pp. 549–555, 2009.

[3] E. Z. Chen and H. Li, "A two-part mixed-effects model for analyzing longitudinal microbiome compositional data," *Bioinformatics*, vol. 32, no. 17, pp. 2611–2617, 2016.

[4] R. J. Little and D. Rubin, "Statistical analysis with missing data," *New York*, 2002.

[5] A. Gelman, H. S. Stern, J. B. Carlin, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

[6] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2018. [Online]. Available: https://www.R-project.org/

[7] NIMBLE Development Team, "NIMBLE: MCMC, particle filtering, and programmable hierarchical modeling," 2019. [Online]. Available: https://cran.r-project.org/package=nimble

[8] P. de Valpine, D. Turek, C. Paciorek, C. Anderson-Bergman, D. Temple Lang, and R. Bodik, "Programming with models: Writing statistical algorithms for general model structures with NIMBLE," *Journal of Computational and Graphical Statistics*, vol. 26, pp. 403–417, 2017.

[9] NIMBLE Development Team, *NIMBLE user manual*, 2018. [Online]. Available: https://r-nimble.org

CHAPTER 3

Results

3.1 Simulation Study

In this section, we describe a simulation study that was conducted to evaluate the performance of the proposed model. We set $N = 100$ and $T = 3$ visits, where $t = 0$ is the baseline for each subject. Two covariates were generated for the analysis, where $x_{i1} \sim N(0, 1)$ and $x_{i2} \sim Bernoulli(0.5)$. For both binary and beta components, same covariates were used. The parameters of the model were set to mimick the real data set used in this study, where it generated approximately similar average zero percentage in the response variable as in real data. The coefficients of fixed-effects of the binary and beta components were set to $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T = (2, 2, 1)^T$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2)^T = (-2, 1, 2)^T$, where β_0 and γ_0 denote the coefficients of the intercept terms, respectively. The random effects of the beta component were generated using multivariate normal distribution, $\boldsymbol{\eta}_i \sim N(\mathbf{0}, \sigma^2\Sigma)$ where $\sigma^2 = 1$ and Σ is a 4×4 AR(1) correlation matrix with $\rho = 0.8$. The response variable y_{it} was thus simulated from model in (1), such that the average overall zero percentage of the response variable to 20%.

In this analysis, the three missing data mechanisms (MCAR, MAR, and MNAR) were used to generate missing data in the response variable. The missing data indicators were simulated for the three cases as follows:

$$R_{itk} \sim Bernoulli(P_{itk}),$$

where $k = 1, 2, 3$ represent each missing data mechanism, $P_{it1} = 0.25$, $P_{it2} = \frac{1}{(1+exp(1+x_{it1}+x_{it2}))}$, and $P_{it3} = \frac{1}{(1+exp(53+x_{it1}+x_{it2}-100y_{it}))}$. Under this setting, the average overall missing data percentage of each missing data mechanism was 25%. This percentage is approximately close to the missing data percentage in real data

set.

The prior distributions for the parameters in the model was set as follows: $\boldsymbol{\beta} \sim N(\mathbf{0}, 10^4 I_3)$, $\boldsymbol{\gamma} \sim N(\mathbf{0}, 10^4 I_3)$, $\phi \sim \text{IG}(10^{-3}, 10^{-3})$, $\sigma^2 \sim \text{IG}(10^{-3}, 10^{-3})$, and $\rho \sim U(0, 1)$. After setting up the priors, 100 simulations were conducted for each missing data mechanism and the true values of the parameters were assigned as the initial values. To acquire the convergence of MCMC samples, the first 1000 iterations of the sampler were discarded and obtained every tenth iteration from 9000 iterations. Therefore, 900 samples were generated from the posterior distributions for each parameter and it was used to compute the posterior summaries.

Tables 1, 2, and 3 show the posterior summaries of each parameter calculated using the results obtained from the three scenarios MCAR, MAR, and MNAR. We reported true value of the parameter (TRUE), posterior median (MEDIAN), posterior mean (MEAN), posterior standard deviation (SD), simulation error (SE), root mean square error of the posterior median (RMSE), and coverage probability (CP) of 95% highest posterior density (HPD) interval (see (15),(16)).

Let $l = (1, \ldots, L)$ denotes the number of simulations conducted in the analysis.

$$\text{MEAN} = \frac{\sum_{l=1}^{L} \theta_l^{mean}}{L} \qquad \text{MEDIAN} = \frac{\sum_{l=1}^{L} \theta_l^{median}}{L} \qquad \text{SD} = \frac{\sum_{l=1}^{L} \theta_l^{sd}}{L} \tag{15}$$

$$\text{SE} = \sqrt{\frac{\sum_{l=1}^{L} (\theta_l^{median} - \text{MEDIAN})^2}{L}} \qquad \text{RMSE} = \sqrt{\frac{\sum_{l=1}^{L} (\theta_l^{median} - \theta^{true})^2}{L}}, \tag{16}$$

where, $\theta_l^{mean}, \theta_l^{median}, \theta_l^{sd}, MEDIAN,$ and θ^{true} represent the posterior mean, posterior median, and posterior standard deviation of l^{th} simulation, overall median of all simulations, and true parameter value respectively.

According to tables 1 and 2, the MEDIAN and MEAN values of the parameters were closer to the true values This can be also seen in the figures 4 and 5,

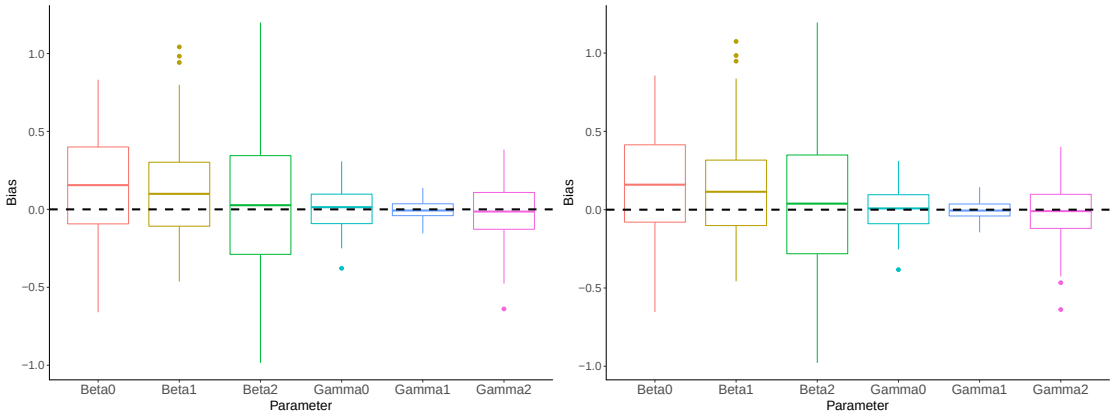
where the bias of coefficient parameters were closer to zero except for ϕ. However, the posterior medians were much closer to the true values compared to the posterior means. Moreover, SD, SE, and RMSE were closer to each other for all the parameters except for ϕ. However, the CP of all the parameters were closer to 95% in both cases.

In table 3, the MEDIAN and MEAN values of parameters were not close to the true parameter values except for parameters β_0 and β_1. The figure 6 also shows a large bias in all the parameters except for β_0 and β_1. The SD, SE, and RMSE values were close to each other for the parametes $\beta_0, \beta_1, \beta_2, \gamma_0,$ and ρ and the CP of the parameters were smaller than 95% except for parameters β_0 and β_1. These results indicate that the proposed model was more adequate if the missing data were MCAR or MAR than MNAR.

Table 1: Posterior summaries of MCAR data with missing data percentage of 25%.

PARAMETERS	TRUE	MEDIAN	MEAN	SD	SE	RMSE	CP
β_0	2.000	2.132	2.144	0.315	0.350	0.325	0.930
β_1	2.000	2.114	2.126	0.303	0.339	0.321	0.940
β_2	1.000	1.026	1.033	0.408	0.437	0.436	0.930
γ_0	-2.000	-1.989	-1.989	0.134	0.140	0.139	0.920
γ_1	1.000	0.995	0.995	0.060	0.059	0.059	0.950
γ_2	2.000	1.984	1.983	0.179	0.191	0.190	0.890
ϕ	50.000	62.898	75.873	44.467	49.416	49.443	0.960
ρ	0.800	0.795	0.792	0.051	0.049	0.049	0.920
σ^2	1.000	0.995	1.005	0.161	0.149	0.149	0.970

The performance of the proposed model was further examined by considering a mis-specified response model. This model was generated by using an unstructured



(a) (b)

Figure 4: Plots of the bias of parameters calculated using posterior median (a) and the bias of parameters calculated using posterior mean (b) of MCAR data.

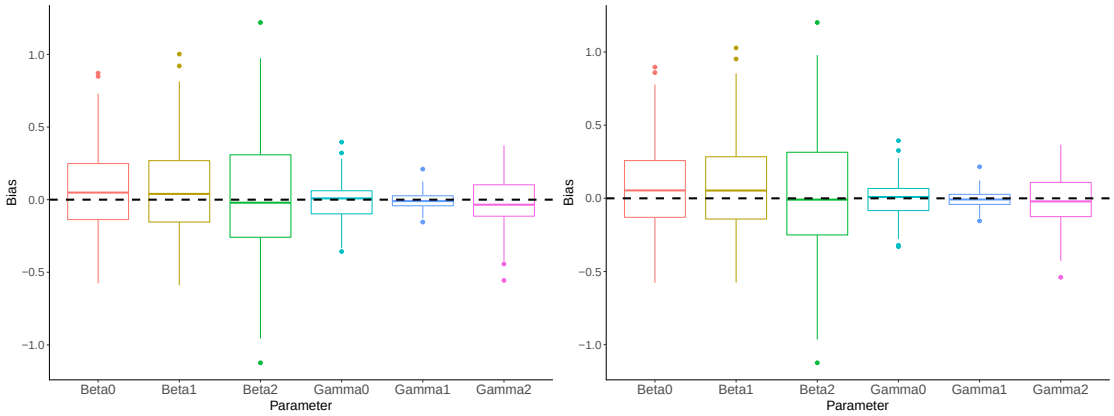
Table 2: Posterior summaries of MAR data with missing data percentage of 25%.

PARAMETERS	TRUE	MEDIAN	MEAN	SD	SE	RMSE	CP
β_0	2.000	2.056	2.066	0.318	0.311	0.306	0.960
β_1	2.000	2.062	2.076	0.327	0.316	0.311	0.950
β_2	1.000	1.032	1.040	0.434	0.444	0.443	0.950
γ_0	-2.000	-2.007	-2.007	0.137	0.138	0.138	0.940
γ_1	1.000	0.994	0.994	0.056	0.057	0.057	0.920
γ_2	2.000	1.981	1.979	0.180	0.187	0.186	0.910
ϕ	50.000	64.954	78.657	45.536	43.494	42.331	0.910
ρ	0.800	0.792	0.790	0.049	0.056	0.055	0.920
σ^2	1.000	0.997	1.007	0.157	0.146	0.146	0.930

correlation matrix Σ of the random effects as follows,

$$\Sigma = \begin{bmatrix} 1.0 & 0.5 & 0.6 & 0.8 \\ 0.5 & 1.0 & 0.4 & 0.9 \\ 0.6 & 0.4 & 1.0 & 0.5 \\ 0.8 & 0.9 & 0.5 & 1.0 \end{bmatrix}$$

Then the response variable y_{it} was generated using the above unstructured correlation matrix. The missing data for y_{it} were simulated using the MAR missing data mechanism with average overall missing data percentage of 25%. The analysis was conducted using the proposed model for 100 simulations using the same



(a) (b)

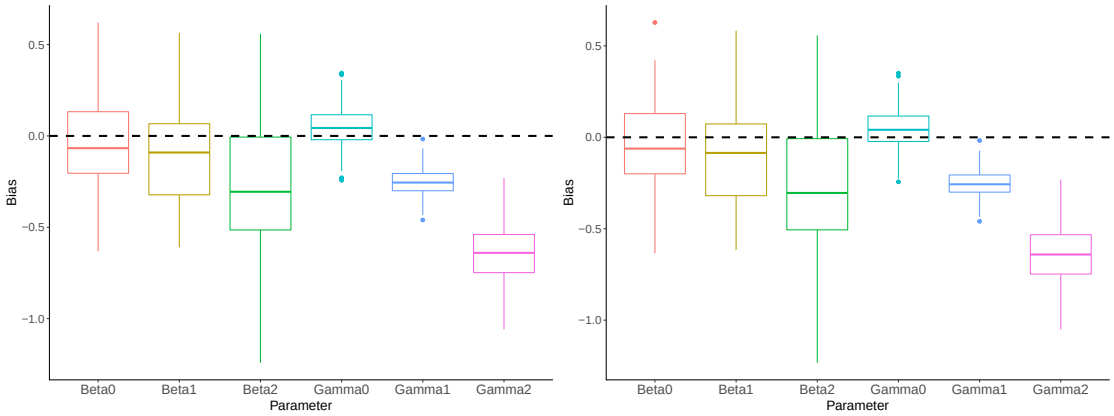
Figure 5: Plots of the bias of parameters calculated using posterior median (a) and the bias of parameters calculated using posterior mean (b) of MAR data.

Table 3: Posterior summaries of MNAR data with missing data percentage of 25%.

PARAMETERS	TRUE	MEDIAN	MEAN	SD	SE	RMSE	CP
β_0	2.000	1.953	1.960	0.262	0.244	0.240	0.960
β_1	2.000	1.910	1.918	0.259	0.276	0.261	0.940
β_2	1.000	0.735	0.739	0.357	0.454	0.368	0.870
γ_0	-2.000	-1.955	-1.955	0.104	0.128	0.120	0.860
γ_1	1.000	0.744	0.744	0.066	0.266	0.073	0.060
γ_2	2.000	1.350	1.350	0.158	0.671	0.167	0.030
ϕ	50.000	29.721	32.510	11.670	24.921	14.883	0.320
ρ	0.800	0.839	0.831	0.067	0.083	0.073	0.850
σ^2	1.000	0.474	0.484	0.112	0.537	0.109	0.040

prior distributions. The first 1000 iterations were discarded in each simulation and every tenth iteration was obtained from 9000 iterations. Moreover, the true values were assigned as the initial values except for the parameter ρ.

Table 4 shows the true values (TRUE), posterior median (EST), SD, SE, RMSE, and CP of each coefficient parameter of fixed effects calculated using the data that were generated from AR(1) structured and unstructured correlation matrices (mis-specified model). When comparing the results of the two models, the



(a) (b)

Figure 6: Plots of the bias of parameters calculated using posterior median (a) and the bias of parameters calculated using posterior mean (b) of MAR data.

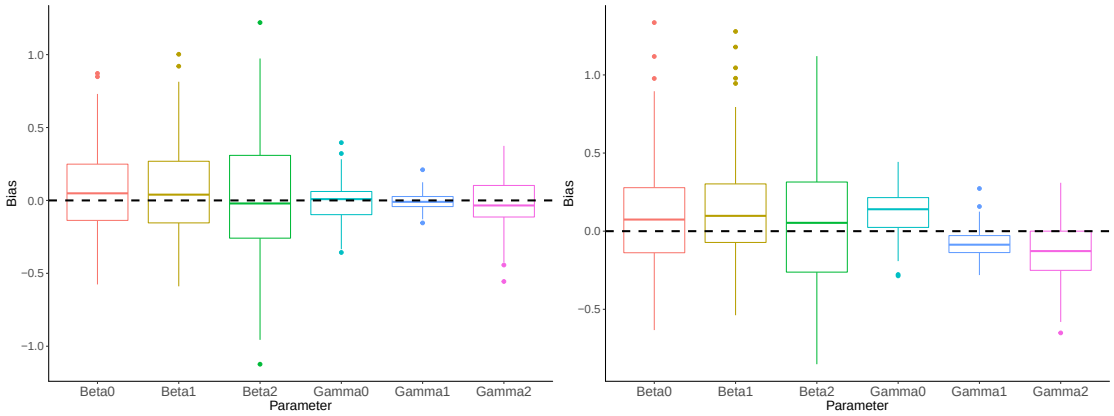
correctly specified model shows better results than the mis-specified model. This also can be clearly seen in figure 7 where the bias of each parameter was closer to zero in the correctly specified model compared to the mis-specified model.

Table 4: Posterior summaries of AR(1) structured and unstructured correlation matrices with MAR data.

		Structured Correlation					Unstructured Correlation				
	TRUE	EST	SD	SE	RMSE	CP	EST	SD	SE	RMSE	CP
β_0	2.000	2.056	0.318	0.311	0.306	0.960	2.119	0.328	0.394	0.378	0.900
β_1	2.000	2.062	0.327	0.316	0.311	0.950	2.147	0.335	0.385	0.358	0.930
β_2	1.000	1.032	0.434	0.444	0.443	0.950	1.030	0.440	0.438	0.437	0.940
γ_0	-2.000	-2.007	0.137	0.138	0.138	0.940	-1.872	0.133	0.198	0.151	0.800
γ_1	1.000	0.994	0.056	0.057	0.057	0.920	0.921	0.063	0.115	0.084	0.680
γ_2	2.000	1.981	0.180	0.187	0.186	0.910	1.879	0.170	0.230	0.196	0.840

3.2 Analysis of Gut Microbiome Composition Data of Preterm Infants

This section depicts the results of the analysis conducted for microbiome composition data of NICU preterm infants. The data of the 68 preterm infants were considered for this analysis using the data of the last four weeks as described in Chapter 2. Therefore, $T = 3$, where $t = 0$ denotes the baseline (week 3) and $t = 1$ to $t = 3$ denotes the follow-up weeks from weeks 4 to 6.



(a) (b)

Figure 7: Plots of the bias of parameters calculated using posterior median in correctly specified model (a) and the bias of the parameters calculated using the posterior median in mis-specified model (b) of MAR data

Table 5 shows the descriptive statistics of the study sample. The median gestational age of the preterm infants was 28 weeks, median birth weight was 1082.5g, approximately 53% of the infants were not treated with antibiotic, and the majority of the infants in this sample were female (approximately 53%). Moreover, the median cumulative weekly stress scores of infants were decreased as the number of weeks in NICU increase.

The same covariates, week (week 3 - reference level), antibiotic usage (1 = yes), infant's gender (1 = female), birth weight, gestational age, and the weekly stress score which is 2 weeks and 1 week prior, used in both binary and beta components of the analysis. The continuous variables, birth weight, gestational age, and stress scores were standardized prior to the analysis.

Out of 21 genera, the Bayesian analysis was conducted for the genus, *'Veillonella'*. The reason behind the selection was that NICU stress exposure had a significant effect on this genus during D'Agata et al., [1] study. To obtain the MCMC samples from the posterior distributions of the parameters, 600,000 iterations were used, and every 500^{th} iteration was taken after removing the first

Table 5: Descriptive statistics of study sample, $n = 68$. Shown as count (%) or median (interquartile range).

Characteristic		
Baby gender		
Male (0)	32	(47.06%)
Female (1)	36	(52.94%)
Antibiotic		
No (0)	36	(52.94%)
Yes (1)	32	(47.06%)
Birth Weight (g)	1082.50	(950.00, 1221.25)
Gestational Age (weeks)	28.35	(26.50, 29.60)
Cumulative Weekly Stress Score		
Week 1	928.50	(879.75, 974.00)
Week 2	918.50	(838.00, 1012.00)
Week 3	881.00	(782.00, 999.00)
Week 4	873.00	(779.75, 977.00)
Week 5	812.00	(774.25, 921.00)
Week 6	802.00	(763.00, 895.00)

100,000 samples due to the high autocorrelation between the MCMC samples. The posterior summaries, posterior median (EST_B) and 95% HPD intervals, were obtained for all the model parameters. The performance of the MCMC samples of parameters were examined using the trace plots and the autocorrelation plots.

Tables 6 and 7 depict posterior summaries of genus *Veillonella* obtained employing the Bayesian approach, where $n = 47$ and $n = 68$ in respective tables. Both tables include, estimates (EST_F) and FDR-adjusted p-values (P-Value) of the parameters obtained from the study D'Agata et al. [1], where they used the frequentist approach with $n = 47$. In both approaches, estimated parameters were considered as statistically significant at the 0.05 significance level. In the Bayesian approach, the estimate is considered as statistically significant if the 95% HPD

interval did not include 0.

According to table 6, in both approaches, the stress two weeks prior (Stress_{t-2}) has a significant effect on the genus *Veillonella* in the beta component. This suggests that when the genus *Veillonella* was present in the stool sample, larger stress scores were associated with the larger relative abundance of the genus. In the Bayesian approach, antibiotic usage was significant in the binary component. This indicates antibiotic usage is associated with a higher probability of *Veillonella* being present in the stool sample. i.e. using antibiotic is 2.147 (exp(0.764)) times the odds to have the genus to be present in the stool sample than non-antibiotic users. The coefficient of week 6 is significant in the beta component in both approaches which shows that sampling week has a significant effect on *Veillonella*.

In table 7 the antibiotic usage in the binary component was not significant in the Bayesian approach. However, it was significant in the beta component where the infants who were treated with antibiotics were associated with the larger relative abundance of *Veillonella* compared to infants who were not treated with an antibiotic. Similar to table 6, the stress two weeks prior collecting the stool sample was significant in the beta component in Bayesian approach. This indicates that the Bayesian approach also agrees that the stress exposure of the preterm infants treated in NICU has a significant effect on the genus *Veillonella*.

List of References

[1] A. L. DAgata, J. Wu, M. K. Welandawe, S. V. Dutra, B. Kane, and M. W. Groer, "Effects of early life NICU stress on the developing gut microbiome," *Developmental Psychobiology*, 2019.

Table 6: Posterior summaries and estimates of binary and beta components of ZIBR model for Bayesian (n = 47) and frequentist (n = 47) approaches for genus *Veillonella*.

Binary Component				
	Bayesian Approach		Frequentist Approach	
	EST_B	95% HPD Interval	EST_F	P-value
Intercept	-1.441	(-2.419, -0.562)	-3.306	0.004
Week 4	-0.129	(-1.232, 0.844)	0.086	0.908
Week 5	-0.002	(-1.034, 1.021)	0.197	0.795
Week 6	0.367	(-0.597, 1.313)	1.208	0.098
Antibiotic	0.764	(0.021, 1.496)	1.971	0.749
Baby gender	-0.015	(-0.795, 0.683)	-0.786	1.000
Birth weight	0.157	(-0.258, 0.544)	-0.254	0.814
Gestational age	0.398	(-0.073, 0.842)	0.910	1.000
Stress_{t-2}	-0.270	(-0.737, 0.201)	-0.598	0.138
Stress_{t-1}	-0.035	(-0.550, 0.483)	0.087	0.825
Beta Component				
	Bayesian Approach		Frequentist Approach	
	EST_B	95% HPD Interval	EST_F	P-value
Intercept	-4.496	(-5.887, -3.184)	-3.385	0.000
Week 4	0.393	(-1.345, 1.524)	0.207	0.510
Week 5	0.372	(-0.869, 1.261)	-0.172	0.606
Week 6	-1.154	(-2.079, -0.384)	-0.693	0.043
Antibiotic	0.541	(-0.435, 1.846)	0.486	0.130
Baby gender	0.463	(-0.407, 1.319)	0.065	0.823
Birth weight	0.019	(-0.481, 0.763)	0.155	0.327
Gestational age	0.338	(-0.151, 1.095)	0.236	0.098
Stress_{t-2}	1.057	(0.391, 1.593)	0.442	0.046
Stress_{t-1}	-0.354	(-0.700, 0.040)	-0.334	0.120

Table 7: Posterior summaries and estimates of binary and beta components of ZIBR model for Bayesian (n = 68) and frequentist (n = 47) approaches for genus *Veillonella*.

Binary Component				
	Bayesian Approach		Frequentist Approach	
	EST_B	95% HPD Interval	EST_F	P-value
Intercept	-1.078	(-1.843, -0.236)	-3.306	0.004
Week 4	-0.013	(-0.852, 0.849)	0.086	0.908
Week 5	0.195	(-0.706, 0.995)	0.197	0.795
Week 6	0.426	(-0.485, 1.292)	1.208	0.098
Antibiotic	0.135	(-0.471, 0.741)	1.971	0.749
Baby gender	-0.078	(-0.739, 0.662)	-0.786	1.000
Birth weight	-0.057	(-0.475, 0.301)	-0.254	0.814
Gestational age	0.054	(-0.426, 0.514)	0.910	1.000
Stress_{t-2}	-0.243	(-0.751, 0.187)	-0.598	0.138
Stress_{t-1}	-0.260	(-0.739, 0.317)	0.087	0.825
Beta Component				
	Bayesian Approach		Frequentist Approach	
	EST_B	95% HPD Interval	EST_F	P-value
Intercept	-4.383	(-5.187, -3.779)	-3.385	0.000
Week 4	0.690	(-0.245, 1.102)	0.207	0.510
Week 5	0.134	(-0.718, 0.838)	-0.172	0.606
Week 6	-1.059	(-1.926, -0.263)	-0.693	0.043
Antibiotic	0.733	(0.192, 1.396)	0.486	0.130
Baby gender	0.545	(-0.497, 1.165)	0.065	0.823
Birth weight	0.401	(-0.273, 0.863)	0.155	0.327
Gestational age	0.425	(-0.090, 0.950)	0.236	0.098
Stress_{t-2}	0.947	(0.665, 1.389)	0.442	0.046
Stress_{t-1}	-0.610	(-1.102, 0.115)	-0.334	0.120

CHAPTER 4

Discussion

In this study, we developed a ZIBR model with mixed effects, which can handle the MAR data using the Bayesian approach. This model can be used to overcome two main challenges, the presence of a large number of zeros in microbiome composition data and intermittent missingness with MAR missing data mechanism that occurs in longitudinal studies.

The performance of the proposed ZIBR model was evaluated by a simulation study (see Section 3.1 in Chapter 3). Estimates of the parameters in the simulation study were closer to true values, except for precision parameter ϕ when the missing data mechanism is MAR. In addition, the coverage probabilities of each parameter were closer to 95%. When the missing data mechanism is MNAR, most of the parameter estimates were not close to true values and the coverage probabilites were also far from 95%. The large bias of the precision parameter ϕ in the beta component highlights deserves future research. Performance of the model was further examined by simulating another data set employing an unstructured correlation matrix for random effects but using the same proposed model. Moreover, the proposed model was applied to real data on gut microbiome composition of NICU preterm infants.

A major limitation faced in this study was conducting the analysis for all the 21 genera. The multiple comparison issue occurs when performing a higher number of tests simultaneously. Thus, we were not able to report the results of the all 21 genera and instead we report the results of one genus. There are several methods that can be used to adjust this issue in a frequentist framework such as Bonferroni correction and false discovery rate (FDR) [1, 2, 3]. In Bayesian framework, one

approach is to use a joint model to model all the genera to solve this issue [4, 5]. Therefore, for future research a multivariate zero inflated model can be used to model all bacterial genera together [6].

The Monte Carlo Expectation Maximization (MCEM) algorithm introduced by Wei and Tanner [7] can be applied to obtain the maximum likelihood estimates (MLE) of parameters for mixed-effect models. In this method, the random effects were treated as unobserved or missing values [8]. This approach can be applied to obtain MLE of parameters of the proposed model. With this approach, one can obtain the fisher's information matrix of parameter estimates and can compare the information gain between the complete case and all case analysis [9]. For future research it is interesting to use the MCEM algorithm to estimate the parameters of the proposed model.

List of References

[1] H. Abdi, "Bonferroni and šidák corrections for multiple comparisons," *Encyclopedia of Measurement and Statistics*, vol. 3, pp. 103–107, 2007.

[2] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[3] Y. Benjamini, D. Yekutieli, *et al.*, "The control of the false discovery rate in multiple testing under dependency," *The Annals of Statistics*, vol. 29, no. 4, pp. 1165–1188, 2001.

[4] S. T. Jensen, I. Erkan, E. S. Arnardottir, D. S. Small, *et al.*, "Bayesian testing of many hypotheses × many genes: A study of sleep apnea," *The Annals of Applied Statistics*, vol. 3, no. 3, pp. 1080–1101, 2009.

[5] K.-A. Do, P. Müller, and F. Tang, "A Bayesian mixture model for differential gene expression," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 54, no. 3, pp. 627–644, 2005.

[6] K. H. Lee, B. A. Coull, A.-B. Moscicki, B. J. Paster, and J. R. Starr, "Bayesian variable selection for multivariate zero-inflated models: Application to microbiome count data," *arXiv Preprint arXiv:1711.00157*, 2017.

[7] G. C. Wei and M. A. Tanner, "A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms," *Journal of the American Statistical Association*, vol. 85, no. 411, pp. 699–704, 1990.

[8] Y. Liu and M. Bottai, "Mixed-effects models for conditional quantiles with longitudinal data," *The International Journal of Biostatistics*, vol. 5, no. 1, 2009.

[9] Q. Chen, J. G. Ibrahim, M.-H. Chen, and P. Senchaudhuri, "Theory and inference for regression models with missing responses and covariates," *Journal of Multivariate Analysis*, vol. 99, no. 6, pp. 1302–1331, 2008.

BIBLIOGRAPHY

Abdi, H., "Bonferroni and šidák corrections for multiple comparisons," *Encyclopedia of Measurement and Statistics*, vol. 3, pp. 103–107, 2007.

Alekseyenko, A. V., Perez-Perez, G. I., De Souza, A., Strober, B., Gao, Z., Bihan, M., Li, K., Methé, B. A., and Blaser, M. J., "Community differentiation of the cutaneous microbiota in psoriasis," *Microbiome*, vol. 1, no. 1, p. 31, 2013.

Bailey, M. T., Dowd, S. E., Parry, N. M., Galley, J. D., Schauer, D. B., and Lyte, M., "Stressor exposure disrupts commensal microbial populations in the intestines and leads to increased colonization by citrobacter rodentium," *Infection and Immunity*, vol. 78, no. 4, pp. 1509–1519, 2010.

Benjamini, Y. and Hochberg, Y., "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

Benjamini, Y., Yekutieli, D., *et al.*, "The control of the false discovery rate in multiple testing under dependency," *The Annals of Statistics*, vol. 29, no. 4, pp. 1165–1188, 2001.

Bull, M. J. and Plummer, N. T., "Part 1: The human gut microbiome in health and disease," *Integrative Medicine: A Clinician's Journal*, vol. 13, no. 6, p. 17, 2014.

Carabotti, M., Scirocco, A., Maselli, M. A., and Severi, C., "The gut-brain axis: Interactions between enteric microbiota, central, and enteric nervous systems," *Annals of Gastroenterology: Quarterly Publication of the Hellenic Society of Gastroenterology*, vol. 28, no. 2, p. 203, 2015.

Chen, E. Z. and Li, H., "A two-part mixed-effects model for analyzing longitudinal microbiome compositional data," *Bioinformatics*, vol. 32, no. 17, pp. 2611–2617, 2016.

Chen, Q., Ibrahim, J. G., Chen, M.-H., and Senchaudhuri, P., "Theory and inference for regression models with missing responses and covariates," *Journal of Multivariate Analysis*, vol. 99, no. 6, pp. 1302–1331, 2008.

Cho, I. and Blaser, M. J., "The human microbiome: At the interface of health and disease," *Nature Reviews Genetics*, vol. 13, no. 4, p. 260, 2012.

Cho, I., Yamanishi, S., Cox, L., Methé, B. A., Zavadil, J., Li, K., Gao, Z., Mahana, D., Raju, K., Teitler, I., *et al.*, "Antibiotics in early life alter the murine colonic microbiome and adiposity," *Nature*, vol. 488, no. 7413, p. 621, 2012.

Cryan, J. F. and Dinan, T. G., "Mind-altering microorganisms: The impact of the gut microbiota on brain and behaviour," *Nature Reviews Neuroscience*, vol. 13, no. 10, p. 701, 2012.

DAgata, A. L., Wu, J., Welandawe, M. K., Dutra, S. V., Kane, B., and Groer, M. W., "Effects of early life NICU stress on the developing gut microbiome," *Developmental Psychobiology*, 2019.

de Valpine, P., Turek, D., Paciorek, C., Anderson-Bergman, C., Temple Lang, D., and Bodik, R., "Programming with models: Writing statistical algorithms for general model structures with NIMBLE," *Journal of Computational and Graphical Statistics*, vol. 26, pp. 403–417, 2017.

Dempster, A. P., Laird, N. M., and Rubin, D. B., "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

Dhariwal, A., Chong, J., Habib, S., King, I. L., Agellon, L. B., and Xia, J., "MicrobiomeAnalyst: A web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data," *Nucleic Acids Research*, vol. 45, no. W1, pp. W180–W188, 2017.

Dimmitt, R. A., Staley, E. M., Chuang, G., Tanner, S. C., Soltau, T. D., and Lorenz, R. G., "The role of postnatal acquisition of the intestinal microbiome in the early development of immune function," *Journal of Pediatric Gastroenterology and Nutrition*, vol. 51, no. 3, p. 262, 2010.

Do, K.-A., Müller, P., and Tang, F., "A Bayesian mixture model for differential gene expression," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 54, no. 3, pp. 627–644, 2005.

Ege, M. J., Mayer, M., Normand, A.-C., Genuneit, J., Cookson, W. O., Braun-Fahrländer, C., Heederik, D., Piarroux, R., and von Mutius, E., "Exposure to environmental microorganisms and childhood asthma," *New England Journal of Medicine*, vol. 364, no. 8, pp. 701–709, 2011.

Field, T., "Alleviating stress in newborn infants in the intensive care unit," *Clinics in Perinatology*, vol. 17, no. 1, pp. 1–9, 1990.

Foster, J. A. and Neufeld, K.-A. M., "Gut-brain axis: How the microbiome influences anxiety and depression," *Trends in Neurosciences*, vol. 36, no. 5, pp. 305 – 312, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0166223613000088

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., and Rubin, D. B., *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

Gonzalez, A., King, A., Robeson II, M. S., Song, S., Shade, A., Metcalf, J. L., and Knight, R., "Characterizing microbial communities through space and time," *Current Opinion in Biotechnology*, vol. 23, no. 3, pp. 431–436, 2012.

Grunau, R. E., Holsti, L., Haley, D. W., Oberlander, T., Weinberg, J., Solimano, A., Whitfield, M. F., Fitzgerald, C., and Yu, W., "Neonatal procedural pain exposure predicts lower cortisol and behavioral reactivity in preterm infants in the NICU," *Pain*, vol. 113, no. 3, pp. 293–300, 2005.

Guarner, F. and Malagelada, J.-R., "Gut flora in health and disease," *The Lancet*, vol. 361, no. 9356, pp. 512–519, 2003.

Ibrahim, J. G., Chen, M.-H., and Lipsitz, S. R., "Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable," *Biometrika*, vol. 88, no. 2, pp. 551–564, 2001.

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., and Herring, A. H., "Missing-data methods for generalized linear models: A comparative review," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 332–346, 2005.

Jensen, S. T., Erkan, I., Arnardottir, E. S., Small, D. S., *et al.*, "Bayesian testing of many hypotheses × many genes: A study of sleep apnea," *The Annals of Applied Statistics*, vol. 3, no. 3, pp. 1080–1101, 2009.

La Rosa, P. S., Zhou, Y., Sodergren, E., Weinstock, G., and Shannon, W. D., "Hypothesis testing of metagenomic data," in *Metagenomics for Microbiology*. Elsevier, 2015, pp. 81–96.

Lee, K. H., Coull, B. A., Moscicki, A.-B., Paster, B. J., and Starr, J. R., "Bayesian variable selection for multivariate zero-inflated models: Application to microbiome count data," *arXiv Preprint arXiv:1711.00157*, 2017.

Lewis, J. D., Chen, E. Z., Baldassano, R. N., Otley, A. R., Griffiths, A. M., Lee, D., Bittinger, K., Bailey, A., Friedman, E. S., Hoffmann, C., *et al.*, "Inflammation, antibiotics, and diet as environmental stressors of the gut microbiome in pediatric crohns disease," *Cell Host & Microbe*, vol. 18, no. 4, pp. 489–500, 2015.

Li, Z., Lee, K., Karagas, M. R., Madan, J. C., Hoen, A. G., OMalley, A. J., and Li, H., "Conditional regression based on a multivariate zero-inflated logistic-normal model for microbiome relative abundance data," *Statistics in Biosciences*, pp. 1–22, 2018.

Little, R. J. and Rubin, D., "Statistical analysis with missing data," *New York*, 2002.

Liu, R. T., "The microbiome as a novel paradigm in studying stress and mental health," *The American Psychologist*, vol. 72, no. 7, pp. 655–667, Oct 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29016169

Liu, Y. and Bottai, M., "Mixed-effects models for conditional quantiles with longitudinal data," *The International Journal of Biostatistics*, vol. 5, no. 1, 2009.

Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R., and Peddada, S. D., "Analysis of composition of microbiomes: A novel method for studying microbial composition," *Microbial Ecology in Health and Disease*, vol. 26, no. 1, p. 27663, 2015.

McMurdie, P. J. and Holmes, S., "phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data," *PloS One*, vol. 8, no. 4, p. e61217, 2013.

Myers, T. A., "Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data," *Communication Methods and Measures*, vol. 5, no. 4, pp. 297–310, 2011.

Newnham, C. A., Inder, T., and Milgrom, J., "Measuring preterm cumulative stressors within the NICU: The neonatal infant stressor scale," *Early Human Development*, vol. 85, no. 9, pp. 549–555, 2009.

NIMBLE Development Team, *NIMBLE user manual*, 2018. [Online]. Available: https://r-nimble.org

NIMBLE Development Team, "NIMBLE: MCMC, particle filtering, and programmable hierarchical modeling," 2019. [Online]. Available: https://cran.r-project.org/package=nimble

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., *et al.*, "A metagenome-wide association study of gut microbiota in type 2 diabetes," *Nature*, vol. 490, no. 7418, p. 55, 2012.

R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2018. [Online]. Available: https://www.R-project.org/

Robins, J. M., Rotnitzky, A., and Zhao, L. P., "Analysis of semiparametric regression models for repeated outcomes in the presence of missing data," *Journal of the American Statistical Association*, vol. 90, no. 429, pp. 106–121, 1995.

Rubin, D. B., *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, 2004, vol. 81.

Tamboli, C., Neut, C., Desreumaux, P., and Colombel, J., "Dysbiosis in inflammatory bowel disease," *Gut*, vol. 53, no. 1, pp. 1–4, 2004.

Tang, Z.-Z. and Chen, G., "Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis," *Biostatistics*, 2018.

Thursby, E. and Juge, N., "Introduction to the human gut microbiota," *Biochemical Journal*, vol. 474, no. 11, pp. 1823–1836, 2017.

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., and Gordon, J. I., "An obesity-associated gut microbiome with increased capacity for energy harvest," *Nature*, vol. 444, no. 7122, p. 1027, 2006.

Voigt, A. Y., Costea, P. I., Kultima, J. R., Li, S. S., Zeller, G., Sunagawa, S., and Bork, P., "Temporal and technical variability of human gut metagenomes," *Genome Biology*, vol. 16, no. 1, p. 73, 2015.

Wang, B., Yao, M., Lv, L., Ling, Z., and Li, L., "The human microbiota in health and disease," *Engineering*, vol. 3, no. 1, pp. 71–82, 2017.

Wei, G. C. and Tanner, M. A., "A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms," *Journal of the American Statistical Association*, vol. 85, no. 411, pp. 699–704, 1990.

WHO, "Preterm birth," http://www.who.int/news-room/fact-sheets/detail/preterm-birth, 2018. [Online]. Available: http://www.who.int/news-room/fact-sheets/detail/preterm-birth

Witt, N., Coynor, S., Edwards, C., and Bradshaw, H., "A guide to pain assessment and management in the neonate," *Current Emergency and Hospital Medicine Reports*, vol. 4, no. 1, pp. 1–10, 2016.

Woodward, L. J., Edgin, J. O., Thompson, D., and Inder, T. E., "Object working memory deficits predicted by early brain injury and development in the preterm infant," *Brain*, vol. 128, no. 11, pp. 2578–2587, 2005.

Xia, F., Chen, J., Fung, W. K., and Li, H., "A logistic normal multinomial regression model for microbiome compositional data analysis," *Biometrics*, vol. 69, no. 4, pp. 1053–1063, 2013.

Xia, Y. and Sun, J., "Hypothesis testing and statistical analysis of microbiome," *Genes & Diseases*, vol. 4, no. 3, pp. 138–148, 2017.

Xu, L., Paterson, A. D., Turpin, W., and Xu, W., "Assessment and selection of competing models for zero-inflated microbiome data," *PloS One*, vol. 10, no. 7, p. e0129606, 2015.

Yin, X., Peng, J., Zhao, L., Yu, Y., Zhang, X., Liu, P., Feng, Q., Hu, Y., and Pang, X., "Structural changes of gut microbiota in a rat non-alcoholic fatty liver disease model treated with a chinese herbal formula," *Systematic and Applied Microbiology*, vol. 36, no. 3, pp. 188–196, 2013.