University of Louisville

ThinkIR: The University of Louisville's Institutional Repository

Electronic Theses and Dissertations

5-2013

Generalized estimating equation based zero-inflated models with application to examining the relationship between dental caries and fluoride exposures.

Sheng Xu
University of Louisville

Follow this and additional works at: <https://ir.library.louisville.edu/etd>

Recommended Citation

Xu, Sheng, "Generalized estimating equation based zero-inflated models with application to examining the relationship between dental caries and fluoride exposures." (2013). *Electronic Theses and Dissertations*. Paper 1606.
<https://doi.org/10.18297/etd/1606>

This Master's Thesis is brought to you for free and open access by ThinkIR: The University of Louisville's Institutional Repository. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of ThinkIR: The University of Louisville's Institutional Repository. This title appears here courtesy of the author, who has retained all other copyrights. For more information, please contact thinkir@louisville.edu.

**GENERALIZED ESTIMATING
EQUATION BASED
ZERO-INFLATED MODELS
WITH APPLICATION TO EXAMINING
THE RELATIONSHIP
BETWEEN DENTAL CARIES AND
FLUORIDE EXPOSURES**

By

Sheng Xu
B.S., Computer Science, Huaqiao University, CHINA, 2007
M.P.A., University of Baltimore, 2010

A Thesis
Submitted to the Faculty of the
School of Public Health and Information Sciences
of the University of Louisville in Partial Fulfillment
of the Requirements for the Degree of

Master of Science

Department of Bioinformatics and Biostatistics
University of Louisville
Louisville, Kentucky

May 2013

Copyright 2013 by Sheng Xu

All rights reserved

**GENERALIZED ESTIMATING
EQUATION BASED
ZERO-INFLATED MODELS
WITH APPLICATION TO EXAMINING
THE RELATIONSHIP
BETWEEN DENTAL CARIES AND
FLUORIDE EXPOSURES**

By

Sheng Xu

B.S., Computer Science, Huaqiao University, CHINA, 2007

M.P.A., University of Baltimore, 2010

A Thesis Approved on

April 22, 2013

by the following Thesis Committee:

Maiying Kong, Thesis advisor

Somnath Datta, Thesis co-advisor

Kristina Zierold

DECICATION

This dissertation is dedicated to my parents

Mr. Yutai Xu

and

Ms. Huiying Lin

Who have given me invaluable educational opportunities.

ACKNOWLEDGMENTS

I would like to express my sincere thanks to my thesis advisor, Dr. Maiying Kong, for her insightful guidance and constant encouragement throughout my research. Her dedication, kindness, and patience have made my endeavor successful and rewarding.

I am very grateful to my co-advisor, Dr. Somnath Datta, for his insightful guidance and support throughout my research. I would like to thank Dr. Kristina Zierold for serving on my thesis committee.

I would like to extend my appreciation to other faculty, staff and graduate students in the Department of Bioinformatics and Biostatistics at the University of Louisville for their help, kindness, and friendship, which make my life at Louisville pleasant and enjoyable.

I would like to thank my aunt, Ms. Lirong Zhao, and my uncle-in-law, Mr. Kai Sun, for their endless love and support. Without them, this thesis would not have been possible.

ABSTRACT

GENERALIZED ESTIMATING EQUATION BASED ZERO-INFLATED MODEL WITH APPLICATION TO EXAMINING THE RELATIONSHIP BETWEEN DENTAL CARIES AND FLUORIDE EXPOSURES

Sheng Xu

April 16, 2013

In the study of dental caries, the number of caries is frequently characterized by over-dispersion and excessive zeros. In addition, the numbers of caries from the same subject are correlated. Zero-Inflated (ZI) regression models, such as ZI-Poisson (ZIP), ZI-Negative Binomial (ZINB), have been developed to account for the excessive zeros in count data. However, the existing zero-inflated models assume that the counts are uncorrelated. On the other hand, Generalized Estimating Equations (GEE) have been developed in the literature to estimate the parameters while accounting for the correlations of observations from the same subject. However, the GEE models incorporating excessive zero counts are not widely available. In this paper, we developed GEE based zero inflated negative binomial model (GEE.ZINB) which account for over-dispersion, excessive zeroes as well as the correlations among the observations from the same subject. We have applied GEE.ZINB, the independent ZINB, and GEE without zero inflation models to examining the association between the dental caries and fluoride exposures using the Iowa fluoride study. We have carried out extensive simulations to examine and compare the performances of the three different methods.

TABLE OF CONTENTS

	PAGE
ACKNOWLEDGMENTS	iv
ABSTRACT	v
LIST OF TABLES.....	ii
LIST OF FIGURES	iii
INTRODUCTION	1
Introduction to longitudinal data analysis.....	1
Generalized Estimating Equations (GEEs).....	2
Iowa Fluoride Study on dental caries	3
GEE BASED ZERO-INFLATED NEGATIVE BINOMIAL MODELS.....	11
Introduction	11
Zero-inflated GEE model	12
DATA ANALYSIS	24
The three possible estimation methods for Iowa Fluoride Study data set	24
Analysis results for Iowa Fluoride Study data set	25
A SIMULATION STUDY	30
A REANALYSIS OF THE DENTAL DATA USING THE BOOTSTRAP METHOD ..	36
DISCUSSION AND FUTURE WORK	38
REFERENCES	39
APPENDIX A.....	41
R code for data adjustment and Figure 2 and Figure 3	41
APPENDIX B.....	45
R code for GEE.ZINB	45
APPENDIX C.....	53
R code for Bootstrap Variance	53

APPENDIX D.....	57
R code for Simulation.....	57
CURRICULUM VITAE.....	62

LIST OF TABLES

TABLE	PAGE
1. The analysis results for Iowa Fluoride study	28
2. The analysis results for the reduce model.....	29
3. Simulation results for highly correlated within-subject observations (Scenario 1).....	33
4. Simulation result for moderately correlated within-subject observations (Scenario 2).....	34
5. Simulation result for uncorrelated within-subject observations (Scenario 3).....	35
6. The analysis results based on bootstrap method for Iowa Fluoride Study data.....	37

LIST OF FIGURES

FIGURE	PAGE
1. Illustration of different types of teeth.....	5
2. The frequencies of teeth with different number of caries	6
3. The frequency of teeth with different number of carries, excluding zero counts.....	7
4. Data structure for Iowa Fluoride Study.....	8

CHAPTER I

INTRODUCTION

1.1 Introduction to longitudinal data analysis

Statistical methods for longitudinal/clustered data have been developed in the past two decades. The longitudinal/clustered data has the features that the observations from the same subject (or cluster) are correlated. Failure to consider the correlations of within-subject observations may result in biased estimates and invalid statistical inferences (Hedeker & Gibbons, 2006; Fitzmaurice, Laird, & Ware, 2011). The techniques developed for longitudinal studies can be widely applied to panel data studies, cohort studies and time series analysis in various fields, such as society, epidemiology, and biology (Hedeker, 2004). The advantages of longitudinal (or clustered) are that they can characterize the relationship between response and covariates at individual level as well as at population level (Hedeker & Gibbons, 2006) .

Many statistical methods have been developed for longitudinal/clustered data. (Hedeker & Gibbons, 2006; Fitzmaurice, Laird, & Ware, 2011). When response variable is continuous, linear mixed effect models are often used. When the response variables are categorical, the generalized linear mixed effect models are used to characterize the subject-specific changes of response relative to covariates. The marginal models, i.e., the

generalized estimating equations (GEEs), have been applied to characterize the relationship between response and the covariates at the population level (Fitzmaurice, Laird, & Ware, 2011; Hedeker & Gibbons, 2006). In this project, our interest is to develop such models for handling clustered longitudinal count data that exhibit the properties of over-dispersion and excessive zero counts.

1.2 Generalized Estimating Equations (GEEs)

In the literature, the generalized estimating equations (GEEs) have been applied to analyzing clustered/longitudinal data. The GEE model is based on the first and second moment of the response variable (Albert, Zeger, & Liang, 1988; Liang & Zeger, 1986). Let us denote the response variable for i^{th} subject as Y_i , where $Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{in_i})^T$. GEE model assumes that the mean of Y_i (say, $EY_i = \mu_i = (\mu_{i1}, \dots, \mu_{in_i})$) and the variance of Y_i , (say V_i) are functions of the covariates, say X_i , where

$$X_i = \begin{pmatrix} x_{i1}^T \\ \vdots \\ x_{in_i}^T \end{pmatrix} = \begin{pmatrix} x_{i11} & \cdots & x_{i1p} \\ \vdots & \vdots & \vdots \\ x_{in_i1} & \cdots & x_{in_ip} \end{pmatrix}.$$

That is, there is a link function g which relates the mean μ_{ij} with the covariate x_{ij} , i.e.

$g(\mu_{ij}) = x_{ij}^T \beta$. Meanwhile, since the observations from the same subject/cluster are

correlated, the correlation matrix of Y_i (say R_i) is directly modeled in the variance matrix.

That is, $Var(Y_i) = V_i = A_i^2 R_i(\alpha) A_i^2$, where A_i is a diagonal matrix with the diagonal

entry as variance of Y_{ij} . One of the commonly used correlation structure for $R_i(\alpha)$ the

compound symmetric correlation structure as follow:

$$R_i(\alpha) = \begin{bmatrix} 1 & \alpha & \alpha & \alpha & \alpha \\ \alpha & 1 & \alpha & \cdots & \alpha \\ \alpha & \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \alpha & \cdots & 1 \end{bmatrix}.$$

Here α is a scale quantity between 0 and 1, and capture the possible correlations between the with-in subject observation structure implies that the correlation structure implies that the correlation between any two within-subject observations is similar. α could be set as fixed, or to be estimated. A GEE model estimates the regression parameter β by solving the following equation:

$$\sum_{i=1}^N \frac{\partial \mu_i^T}{\partial \beta} V_i^{-1} (Y_i - \mu_i) = 0.$$

The estimation procedure has been implemented in several packages of the statistical analysis software R (R Development Core Team, 2013). Examples include the function `gee` under the R-package “`gee`” (Carey, Lumley, & Ripley, 1998) and the function `geeglm` under the R- package “`geepack`” (Jun & Halekoh, 2012). The functions can easily incorporate different correlation matrices into the GEE models. We have applied the GEE model to analyze a dental data set in Chapter 3. However, the standard GEE model is not sufficient to describe count data which has excessive zeros. One such example is the counts of dental caries that has extremely large amount of zeros. This example is presented in the following sub-section.

1.3 Iowa Fluoride Study on dental caries

The Iowa Fluoride Study (IFS) is a longitudinal study of children designed to quantify fluoride exposures from both dietary and non-dietary sources and to associate

longitudinal fluoride exposures with dental fluorosis (spots on teeth) and dental caries. Mothers of newborns were recruited from 8 Iowa hospital postpartum units between 1992 and 1995. These hospitals were responsible for the large majority of all births in the area that they served, with a combined total of about 8,000 births per year, or approximately 20 percent of all births in Iowa (Iowa Fluoride Study , 2013). In this project, we are interested in investigating whether the numbers of dental caries in children are associated fluoride exposure at a pre-specified time point. Dental caries, also known as tooth decay or a cavity, is an infection, bacterial in origin that causes demineralization and destruction of the hard tissues (Wikipedia encyclopedia). In the Iowa Fluoride Study, the numbers of carries for each surface of individual tooth within a subject were recorded. In this study, there were at most two carries in each surface of a tooth. Some teeth have 4 surfaces, and some have 5 surfaces, which implies that the number of carries on a tooth is at most 10 (See Figure 1 (Mayo foundation for Medical Education and Research, 2012)).

Figure 1. Illustration of different types of teeth.

The frequencies of teeth with different numbers of caries for the entire study subjects are shown in the Figure 2. From the Figure 2, it is clear that the number of zero counts are excessive large. The frequencies of teeth excluding zero counts of caries are shown in Figure 3, which indicates that the number of caries, excluding excessive zeros, is more like truncated Poisson or Negative Binomial distributions.

Figure 2. The frequencies of teeth with different number of caries indicate excessive zero counts for caries.

Figure 3. The frequency of teeth with different number of carries, excluding zero counts.

In the literature, zero-inflated models have been developed to model the cases where excessive zeros exist (Wan, Hua, & Xin, 2012). However, the zero-inflated models do not account for the correlations among the observations from the same subject/cluster.

In the current project, the Iowa Fluoride Studies involve three data sets, which are illustrated in Figure 4. Data set 1 includes the subject's gender, social economic status,

and race. Data set 2 includes different treatments: daily soda pop intake (AUCSodaOz0_5yrs), average of all tooth brushing frequencies reported for the past 6 months (DentalVisitPast6moAvg), the average of times a professional dental fluoride treatment for the preceding period (FluorideTreatment6moAvg), and the average home tap water fluoride level for all the return questionnaires (HomeFluorideppmAvg). Data set 3 includes the response variables, the number of caries at each surface for each tooth within a subject. All data sets are linked by the identification number (SID). The response variable is summarized as the number of caries for each tooth.

Figure 4. Data structure for the Iowa Fluoride Study.

Most previous work in analyzing these data sets used simpler odds ratio type calculations to assess various potential risk factors. In some model building attempts, a (multivariable/stepwise, etc.) logistic regression approach was taken (Levy, Warren, Broffitt, Hillis, & Kanellis, 2003; Marshall, Broffitt, Eichenberger-Gilmore, Warren, Cunningham, & Levy, 2005; Marshall, Eichenberger-Gilmore, Larson, Warren, & Levy, 2007a; Marshall, Eichenberger-Gilmore, Broffitt, Warren, & Levy, 2007b; Hong, Levy, Warren, & Broffitt, 2009; Iida & Kumar, 2009). While this is an extremely popular statistical technique in medical research in general, the analysis often oversimplifies facts. In addition, dichotomizing the outcome (e.g., caries versus non caries) may lead to inefficient use of the count data (e.g., number of carious surfaces or scores). Broffitt et al. (2007) and Chankanka (2010) applied the generalized linear model techniques in analyzing caries data as count data regression; they both settled on the negative binomial model.

1.4 Zero-inflated regression models

The numbers of carries in the Iowa fluoride study have extremely large proportion of zeros (see Figure 2), which is beyond what a standard Poisson or negative binomial models can describe. In the literature, zero-inflated regression model has been developed to model count data with excessive zeros (Zeileis, Kleiber, & Jackman, 2008). Zero-inflated regression models consist of two regression models: a logistic or probit regression model and a count model. The logistic regression models the probability of excess zeros in terms of available covariates. The count model relates the mean of count with available covariates using the framework of generalized linear model when the

response is not from the distribution degenerated at zero part. The standard zero-inflated model has been implemented as a function *zeroinfl* in R-package “pscl” (Jackman, Tahk, Zeileis, Maimone, & Fearon, 2012). In the standard zero-inflated model, the counts are assumed to be independent. However the counts of caries for the teeth from a subject are correlated. In Chapter 2, we extended the zero-inflated model to model correlated counts, which is considered as the GEE based zero-inflated model. The algorithm has been implemented using R; the code is placed in the appendix. In Chapter 3, we applied this method to analyzing the dental caries data. In Chapter 4, simulations are carried out to compare the performance of the proposed model and the two other models. In Chapter 5, we reanalyzed the dental data using resampling (bootstrap) based variance estimates. The last chapter is devoted to a discussion and potential future work.

CHAPTER II

GEE BASED ZERO-INFLATED NEGATIVE BINOMIAL MODELS

2.1 Introduction

Negative binomial and Poisson models have been applied widely to count data. In the case when the counts for zeros are above and beyond the number of sampling zeros expected by the negative binomial distribution or the Poisson distribution, a degenerated distribution at zero is introduced to account for the extra zeros, which are often called the zero-inflated models. In the Iowa Fluoride Study data set, the number of caries in each tooth within a subject has been recorded, and the numbers of caries for all teeth within a subject are most likely correlated. Without loss of generality, let us denote Y_{ij} as the number of caries for the i^{th} tooth within i^{th} subject. Let

denote $Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{in_i})^T$ ($i = 1, \dots, N$), and the associated covariates as

$$X_i = \begin{pmatrix} x_{i1}^T \\ \dots \\ x_{in_i}^T \end{pmatrix} = \begin{pmatrix} x_{i11} & \dots & x_{i1p} \\ \dots & \dots & \dots \\ x_{in_i1} & \dots & x_{in_ip} \end{pmatrix}. \quad (1)$$

The distribution of Y_{ij} follows a mixture of a degenerated distribution at zero with mixing probability of p_{ij} and a negative binominal (NB) distribution or Poisson distribution with mean λ_{ij} with mixing probability $1 - p_{ij}$, and $\log(\lambda_{ij}) = x_{ij}^T \beta$. In general, p_{ij} is modeled

by a logit function, such as $\text{logit}(p_{ij}) = z_{ij}^T \gamma$, where z_{ij} could be different from x_{ij} or a subset of x_{ij} ($i = 1, \dots, N; j = 1, \dots, n_i$).

2.2 Zero-inflated GEE model

Suppose that W_{ij} follows a NB distribution with mean λ_{ij} and a shape parameter τ . The distribution function for W_{ij} can be written as

$$P(W_{ij} = w_{ij}) = f_{NB}(w_{ij}|\lambda_{ij}, \tau) = \frac{\Gamma(w_{ij} + \frac{1}{\tau})}{w_{ij}! \Gamma(\frac{1}{\tau})} \left(\frac{1}{1 + \tau \lambda_{ij}}\right)^{\frac{1}{\tau}} \left(\frac{\tau \lambda_{ij}}{1 + \tau \lambda_{ij}}\right)^{w_{ij}}, \quad (2)$$

where $w_{ij} = 0, 1, \dots; \tau$ ($\tau > 0$) is a shape parameter that quantifies the amount of over-dispersion. The mean and variance of W_{ij} are given by $E(W_{ij}|\lambda_{ij}, \tau) = \lambda_{ij}$, and $Var(W_{ij}|\lambda_{ij}, \tau) = \lambda_{ij} + \tau \lambda_{ij}^2$. Unless $\tau = 0$, the variance is always larger than the mean λ_{ij} . Thus, NB model adds a quadratic term $\tau \lambda_{ij}^2$ to the variance of Poisson to account for the extra-Poisson variation or over-dispersion. τ is known as the dispersion or shape parameter (Wan, Hua, & Xin, 2012). The NB distribution gets closer to the Poisson distribution if τ becomes smaller, i.e., $f_{NB}(w_{ij}) \rightarrow f_P(w_{ij})$ as $\tau \rightarrow 0$. Thus, the larger value of τ is, the more variability is in the data set which is over and beyond that a Poisson distribution can describe.

The zero-inflated NB models the count data where zero counts are beyond that a NB distribution can describe. Let us denote the response variable as Y_{ij} , which can be written as the mixture of a degenerated distribution at zero, and a NB model random variable W_{ij} . Thus, the distribution of the response variable Y_{ij} can be written as:

$$P(Y_{ij} = y) = \begin{cases} p_{ij} + (1 - p_{ij}) \Pr(W_{ij} = 0) & \text{if } y = 0 \\ (1 - p_{ij}) \Pr(W_{ij} = y) & \text{if } y \geq 1 \end{cases} \quad (3a)$$

$$= \begin{cases} p_{ij} + (1 - p_{ij}) \left(\frac{1}{1 + \tau \lambda_{ij}} \right)^{\frac{1}{\tau}} & \text{if } y = 0 \\ (1 - p_{ij}) \frac{\Gamma(y + \frac{1}{\tau})}{y! \Gamma(\frac{1}{\tau})} \left(\frac{1}{1 + \tau \lambda_{ij}} \right)^{\frac{1}{\tau}} \left(\frac{\tau \lambda_{ij}}{1 + \tau \lambda_{ij}} \right)^y & \text{if } y \geq 1. \end{cases} \quad (3b)$$

The expected value for Y_{ij} is

$$E\{Y_{ij}\} = \sum_{y=0}^{\infty} y P(Y_{ij} = y) = (1 - p_{ij}) \sum_{y=0}^{\infty} y P(W_{ij} = y) = (1 - p_{ij}) \lambda_{ij}. \quad (4)$$

The variance for Y_{ij} is

$$\begin{aligned} \text{Var}\{Y_{ij}\} &= \sum_{y=0}^{\infty} (y - E\{Y_{ij}\})^2 P(Y_{ij} = y) = \sum_{y=0}^{\infty} (y - (1 - p_{ij}) \lambda_{ij})^2 P(Y_{ij} = y) \\ &= p_{ij} (1 - p_{ij})^2 \lambda_{ij}^2 + \sum_{y=0}^{\infty} p_{ij} (1 - p_{ij})^2 (1 - p_{ij}) P(W_{ij} = y) \\ &= p_{ij} (1 - p_{ij})^2 \lambda_{ij}^2 + (1 - p_{ij}) (\lambda_{ij} + \tau \lambda_{ij}^2 + p_{ij}^2 \lambda_{ij}^2) \\ &= (1 - p_{ij}) \lambda_{ij} (1 + \tau \lambda_{ij} + p_{ij} \lambda_{ij}) \end{aligned} \quad (5)$$

When applying the zero-inflated negative binomial model, we must model both p_{ij} and λ_{ij} as functions of explanatory variables. The log link function is used to relate λ_{ij} to the explanatory variables (say, x_{ij}), and the logit link function is used to

relate p_{ij} to the explanatory variables (say, z_{ij}). The predictors (say, x_{ij}) for λ_{ij} may be different from the predictors (say, z_{ij}) for p_{ij} . Without loss of generality, let assume that $\log(\lambda_{ij}) = x_{ij}^T \beta$ and $\text{logit}(p_{ij}) = z_{ij}^T \gamma$ ($i = 1, 2, \dots, N; j = 1, \dots, n_i$). Thus, the mean of Y_{ij} , i.e., $E\{Y_{ij}\} = \mu_{ij} = (1 - p_{ij})\lambda_{ij}$, depends on the parameters β and γ . To account for the correlations for the observations within the same subject (or cluster), we may introduce the correlation matrix, say $R_i(\alpha)$ for i^{th} subject (or cluster), and estimate β and γ by applying the generalized estimating equations (GEE) (Hall & Zhang, 2004):

$$\sum_{i=1}^N \begin{pmatrix} \frac{\partial \mu_i^T}{\partial \beta} \\ \frac{\partial \mu_i^T}{\partial \gamma} \end{pmatrix} V_i^{-1}(\alpha, \beta, \gamma, \tau)(Y_i - \mu_i) = 0. \quad (6)$$

Here $Y_i = (Y_{i1}, \dots, Y_{in_i})^T$, $\mu_i = (\mu_{i1}, \dots, \mu_{in_i})^T$, $V_i(\alpha, \beta, \gamma, \tau) = \text{Var}(Y_i) = A_i^{\frac{1}{2}} R_i(\alpha) A_i^{\frac{1}{2}}$ with $A_i = \text{Diag}\{\text{Var}(Y_{ij}) | j=1, \dots, n_i\}$. This direct application of GEE to the clustered zero-inflated models may not be identifiable because β and γ are typically confounded in equation (6) (Hall & Zhang, 2004). In the following, we estimate β and γ in two separate equations by introducing latent variables u_{ij} ($i = 1, \dots, N; j = 1, \dots, n_i$) (Hall & Zhang, 2004).

Let $u_{ij} = 0$ if $Y_{ij} \sim f_{NB}(\lambda_{ij}, \tau)$ and $u_{ij} = 1$ if Y_{ij} is from zero-degenerated distribution. Thus, $\Pr(u_{ij} = 1) = p_{ij}$ with $\text{logit}(p_{ij}) = z_{ij}^T \gamma$. The GEE for γ can be written as:

$$\sum_{i=1}^N \frac{\partial p_i^T}{\partial \gamma} \{V_{\gamma i}\}^{-1} (u_i - p_i) = 0. \quad (7)$$

Here $p_i^T = (p_{i1}, \dots, p_{in_i})$, and p_{ij} is determined by $\text{logit}(p_{ij}) = \log \frac{p_{ij}}{1-p_{ij}} = z_{ij}^T \gamma$, which implies that $p_{ij}(\gamma) = \frac{\exp(z_{ij}^T \gamma)}{1 + \exp(z_{ij}^T \gamma)}$ and $\frac{\partial p_{ij}}{\partial \gamma^T} = z_{ij} \frac{\exp(z_{ij}^T \gamma)}{(1 + \exp(z_{ij}^T \gamma))^2} \cdot V_{\gamma i} = A_i^{\frac{1}{2}} R_{1i}(\alpha_1) A_i^{\frac{1}{2}}$ is the variance matrix for u_i , where $A_i = \text{Diag}\{p_{i1}(1 - p_{i1}), p_{i2}(1 - p_{i2}), \dots, p_{in_i}(1 - p_{in_i})\}$ with the j^{th} entry being the variance of u_{ij} , where $u_i = (u_{i1}, \dots, u_{in_i})^T$.

Similarly, the generalized estimating equation for β can be written as

$$\sum_{i=1}^N \frac{\partial \lambda_i^T}{\partial \beta} \{V_{\beta i}\}^{-1} \text{Diag}(1 - u_i)(y_i - \lambda_i) = 0. \quad (8)$$

Here $\lambda_i^T = (\lambda_{i1}, \dots, \lambda_{in_i})$, and λ_{ij} is determined by $\log(\lambda_{ij}) = x_{ij}^T \beta$, which implies that $\lambda_{ij} = \exp(x_{ij}^T \beta)$ and $\frac{\partial \lambda_{ij}}{\partial \beta} = x_{ij} \exp(x_{ij}^T \beta)$. $V_{\beta i} = D_i^{1/2} R_{2i}(\alpha_2) D_i^{1/2}$ is the variance matrix for Y_i when Y_i is from NB, where $D_i = \text{Diag}\{\lambda_{i1}(1 + \tau \lambda_{i1}), \lambda_{i2}(1 + \tau \lambda_{i2}), \dots, \lambda_{in_i}(1 + \tau \lambda_{in_i})\}$ and $R_{2i}(\alpha_2)$ is the correlation matrix for Y_i when Y_i is from NB. The diagonal matrix $\text{Diag}\{1 - u_i\} = \text{Diag}(1 - u_{i1}, \dots, 1 - u_{in_i})$, and the matrix $\text{Diag}\{1 - u_i\}$ in equation (8) indicates that only the Y_{ij} from NB distribution (*i.e.*, $u_{ij} = 0$) contributed the estimating equation (8), and the Y_i form zero-degenerated distribution (*i.e.*, $u_{ij} = 1$) does not contribute to equation (8). Given $u_{ij} (i = 1, \dots, N; j = 1, \dots, n_i)$, β and γ usually are estimated by the Fisher-scoring method, which is an iterative algorithm for solving the estimating equations such as (7) and (8). However, since $u_{ij} (i = 1, \dots, N; j = 1, \dots, n_i)$ are latent variables, u_{ij} needs to be estimated at each iteration. Hall and Zhang (2004) provided the estimation for u_{ij} using the Expectation Maximization (EM) algorithm under the

assumption that all observations are independent. In fact, the estimation of u_{ij} can be considered as the posterior mean of u_{ij} given y, γ, β and τ . Suppose that in the b^{th} iteration, the estimate for γ, β , and τ are denoted as $\gamma^{(b)}, \beta^{(b)}$, and $\tau^{(b)}$, u_{ij} can be updated by

$$\begin{aligned}
u_{ij}^{(b)} &= \Pr(u_{ij} = 1 | y, \gamma^{(b)}, \beta^{(b)}, \tau^{(b)}) = \left\{ \frac{\Pr(u_{ij} = 1, y_{ij} = 0 | \gamma^{(b)}, \beta^{(b)}, \tau^{(b)})}{\Pr(y_{ij} = 0 | \gamma^{(b)}, \beta^{(b)}, \tau^{(b)})} \right\} 1_{\{y_{ij}=0\}} \\
&= \left\{ \frac{p_{ij}^{(b)}}{p_{ij}^{(b)} + (1 - p_{ij}^{(b)}) f_{NB}(0 | \lambda_{ij}^{(b)}, \tau^{(b)})} \right\} 1_{\{y_{ij}=0\}} \\
&= \left\{ 1 + \frac{(1 - p_{ij}^{(b)}) f_{NB}(0 | \lambda_{ij}^{(b)}, \tau^{(b)})}{p_{ij}(\gamma^{(b)})} \right\}^{-1} 1_{\{y_{ij}=0\}} \\
&= \left\{ 1 + \frac{(1 - p_{ij}^{(b)}) \left(\frac{1}{1 + \tau^{(b)} \lambda_{ij}^{(b)}} \right)^{\frac{1}{\tau^{(b)}}}}{p_{ij}^{(b)}} \right\}^{-1} 1_{\{y_{ij}=0\}} \tag{9}
\end{aligned}$$

Here $p_{ij}^{(b)} = \frac{\exp(z_{ij}^T \gamma^{(b)})}{1 + \exp(z_{ij}^T \gamma^{(b)})}$, and $\lambda_{ij}^{(b)} = \exp(x_{ij}^T \beta^{(b)})$. Thus, γ and β can be updated by the following iterated formula:

$$\gamma^{(b+1)} = \gamma^{(b)} + \left\{ \sum_{i=1}^N \frac{\partial p_i^T}{\partial \gamma} \{V_{\gamma i}\}^{-1} \frac{\partial p_i}{\partial \gamma^T} \right\}^{-1} \sum_{i=1}^N \frac{\partial p_i^T}{\partial \gamma} \{V_{\gamma i}\}^{-1} (u_i^{(b)} - p_i) |_{(\gamma, \alpha_1) = (\gamma^{(b)}, \alpha_1^{(b)})} \tag{10}$$

and

$$\beta^{(b+1)} = \beta^{(b)} + \left\{ \sum_{i=1}^N \frac{\partial \lambda_i^T}{\partial \beta} \{V_{\beta i}\}^{-1} \text{Diag}(1 - u^{(b)}) \frac{\partial \lambda_i}{\partial \beta^T} \right\}^{-1} S_{\beta}, \tag{11}$$

where $S_\beta = \sum_{i=1}^N \frac{\partial \lambda_i^T}{\partial \beta} \{V_{\beta i}\}^{-1} \text{Diag}(1 - u^{(b)}) (y_i - \lambda_i)$, and (β, τ, α_2) in equation (11) is replaced by $(\beta^{(b)}, \tau^{(b)}, \alpha_2^{(b)})$.

The parameter τ , is not related to the mean function but only related to the variance function when Y_{ij} is from NB. To estimate τ , let us set $\epsilon_{ij}^2(\beta) = (y_{ij} - \lambda_{ij}(\beta))^2$, then we have $E(\epsilon_{ij}^2(\beta)) = \lambda_{ij}(1 + \tau \lambda_{ij}) = v_{ij}$. Let $\epsilon_i^2(\beta) = (\epsilon_{i1}^2(\beta), \dots, \epsilon_{in_i}^2(\beta))^T$ and $E(\epsilon_i^2(\beta)) = v_i(\tau) = (v_{i1}, \dots, v_{in_i})^T$ when Y_i is from NB. Since τ only involves the variance of Y_{ij} , we propose estimating τ by solving the following equations:

$$\sum_{i=1}^N \Gamma_i^T H_i (\epsilon_i^2(\beta) - v_i(\tau)) = 0. \quad (12)$$

Here $\Gamma_i = \left(\frac{\partial(v_i(\tau))}{\partial \tau^T} \right) = \left(\frac{\partial v_i(\tau)}{\partial \tau^T} \right) = \begin{pmatrix} \lambda_{i1}^2 \\ \dots \\ \lambda_{in_i}^2 \end{pmatrix}$ and $H_i = \text{Diag}((1 - u_{i1})^2, \dots, (1 - u_{in_i})^2)$.

Equation (12) specifies the following estimating equation for τ :

$$\sum_{i=1}^N \sum_{j=1}^{n_i} \lambda_{ij}^2 (1 - u_{ij})^2 (\epsilon_{ij}^2(\beta) - v_{ij}(\tau)) = 0. \quad (13)$$

Given β and γ , one can solve for τ from equation (13),

$$\tau = \frac{\sum_{i=1}^N \sum_{j=1}^{n_i} \lambda_{ij}^2 (1 - u_{ij})^2 (\epsilon_{ij}^2(\beta) - \lambda_{ij})}{\sum_{i=1}^N \sum_{j=1}^{n_i} (1 - u_{ij})^2 \lambda_{ij}^4} = \frac{\sum_{i=1}^N \sum_{j=1}^{n_i} \lambda_{ij}^2 (1 - u_{ij})^2 ((y_{ij} - \lambda_{ij})^2 - \lambda_{ij})}{\sum_{i=1}^N \sum_{j=1}^{n_i} (1 - u_{ij})^2 \lambda_{ij}^4}. \quad (14)$$

Next we need to estimate α_1 and α_2 . To estimate α_1 , let us set

$$U_{ist}(\gamma) = \frac{(u_{is} - p_{is})(u_{it} - p_{it})}{\sqrt{p_{is}(1 - p_{is})p_{it}(1 - p_{it})}} \quad (15)$$

which has the expected value of ρ_{ist} . Let us denote $U_{\gamma i} = (U_{i12}, U_{i13}, \dots, U_{in_i-1, n_i})^T$,

and $\rho_{\gamma i}(\alpha_1) = E\{U_{\gamma i}\} = (\rho_{i12}, \rho_{i13}, \dots, \rho_{in_i-1, n_i})^T$. Hence α_1 can be estimated by the following equation:

$$\sum_{i=1}^N E_{\gamma i}^T W_{\gamma i}^{-1} (U_{\gamma i} - \rho_{\gamma i}(\alpha_1)) = 0, \quad (16)$$

where $E_{\gamma i} = \frac{\partial \rho_{\gamma i}(\alpha_1)}{\partial \alpha_1^T}$, and $W_{\gamma i} \cong \text{Cov}(U_{\gamma i})$. In the original paper on GEE, Liang and

Zeger (1986) suggested letting $W_{\gamma i}$ be the identity matrix, while (Prentice, 1988)

suggested letting $W_{\gamma i}$ be a diagonal matrix with the approximate variance of $U_{\gamma i}$ along the diagonal. In case that $W_{\gamma i}$ being the identity matrix, and $R_{\gamma i}(\alpha_1)$ is the symmetric

compound structure, one could obtain the estimate α_1 as follows:

$$\hat{\alpha}_1 = \frac{1}{N^*} \sum_{i=1}^N \sum_{s < t} \frac{(u_{is} - p_{is})(u_{it} - p_{it})}{\sqrt{p_{is}(1 - p_{is})p_{it}(1 - p_{it})}} \quad (17)$$

where $N^* = \sum_{i=1}^N \frac{n_i(n_i-1)}{2}$.

An alternative estimate for α_1 is

$$\hat{\alpha}_1 = \frac{\frac{1}{N^*} \sum_{i=1}^N \sum_{s < t} \frac{(u_{is} - p_{is})(u_{it} - p_{it})}{\sqrt{p_{is}(1 - p_{is})p_{it}(1 - p_{it})}}}{\frac{1}{N_{tot}} \sum_{i=1}^N \sum_{j=1}^{n_i} \frac{(u_{ij} - p_{ij})^2}{p_{ij}(1 - p_{ij})}}, \quad (18)$$

where $N_{tot} = \sum_{i=1}^N n_i$.

To estimate α_2 , let us denote

$$U_{ist}(\beta) = \frac{(y_{is} - \lambda_{is})(y_{it} - \lambda_{it})}{\sqrt{\lambda_{is}(1 + \tau\lambda_{is})\lambda_{it}(1 + \tau\lambda_{it})}} \quad (19)$$

and $U_{\beta i} = (U_{i12}, U_{i13}, \dots, U_{in_i-1, n_i})^T$, which has expected value of $\rho_{\beta i}(\alpha_2)$,

where $\rho_{\beta i}(\alpha_2) = E\{U_{\beta i}\} = (\rho_{i12}, \rho_{i13}, \dots, \rho_{in_i-1, n_i})^T$ when Y_i is from NB. Thus, α_2 can

be estimated by the following equation:

$$\sum_{i=1}^N E_{\beta i}^T W_{\beta i}^{-1} H_{\beta i} (U_{\beta i} - \rho_{\beta i}(\alpha_2)) = 0. \quad (20)$$

Here $E_{\beta i} = \frac{\partial \rho_{\beta i}(\alpha_2)}{\partial \alpha_2^T}$, $W_{\beta i} \cong \text{Cov}(U_{\beta i})$, and

$H_{\beta i} = \text{Diag}\{(1 - u_{i1})(1 - u_{i2}), \dots, (1 - u_{in_i-1})(1 - u_{in_i})\}$. In the case that $W_{\beta i}$ is the identity matrix and $R_{\beta i}(\alpha_2)$ is the symmetric compound structure, α_2 could be obtained as follows:

$$\hat{\alpha}_2 = \frac{1}{N^*} \sum_{i=1}^N \sum_{s < t} \frac{(1 - u_{is})(1 - u_{it})(y_{is} - \lambda_{is})(y_{it} - \lambda_{it})}{\phi \sqrt{\lambda_{is}(1 + \tau\lambda_{is})\lambda_{it}(1 + \tau\lambda_{it})}}. \quad (21)$$

Here $N^* = \sum_{i=1}^N \sum_{s < t} (1 - u_{is})(1 - u_{it})$. An alternative estimate for α_2 is

$$\hat{\alpha}_2 = \frac{1}{N^*} \sum_{i=1}^N \sum_{s < t} \frac{\frac{(1-u_{is})(1-u_{it})(y_{is}-\lambda_{is})(y_{it}-\lambda_{it})}{\sqrt{\lambda_{is}(1+\tau\lambda_{is})\lambda_{it}(1+\tau\lambda_{it})}}}{\frac{1}{N_{tot}} \sum_{i=1}^N \sum_{j=1}^{n_i} \frac{(1-u_{ij})^2(y_{ij}-\lambda_{ij})^2}{\lambda_{ij}(1+\tau\lambda_{ij})}}, \quad (22)$$

where $N_{tot} = \sum_{i=1}^N \sum_{j=1}^{n_i} (1-u_{ij})^2$. In the current project, the alternative estimates for α_1 and α_2 were used.

To obtain the final estimate for $\beta, \gamma, \tau, \alpha_1$ and α_2 , an iterative method is required to iterate between estimating β and γ (given the current estimate of τ, α_1 and α_2) as the solution of equation (7) and (8), and estimating τ, α_1 and α_2 (given the current estimate of β and γ) as the solution of (13), (16) and (20) until convergence. The iterative method can be implemented as the following steps:

Step 1: Given initial values for the parameter estimates of $\beta, \gamma, \tau, \alpha_1$ and α_2 , denoted as $\hat{\beta}^{(0)}, \hat{\gamma}^{(0)}, \hat{\tau}^{(0)}, \hat{\alpha}_1^{(0)}$, and $\hat{\alpha}_2^{(0)}$; set $b = 0$.

Step 2: Update the latent variable u_{ij} ($i = 1, \dots, N; j = 1, \dots, n_i$), as defined in equation (9), in the b^{th} iteration:

$$\hat{u}_{ij}^{(b)} = \left\{ 1 + \frac{(1-\hat{p}_{ij}^{(b)}) \left(\frac{1}{1+\hat{\tau}^{(b)}\hat{\lambda}_{uj}^{(b)}} \right)^{\frac{1}{\hat{\tau}^{(b)}}}}{\hat{p}_{ij}^{(b)}} \right\}^{-1} 1\{y_{ij} = 0\}.$$

Step 3: Given $\hat{\beta}^{(b)}, \hat{\gamma}^{(b)}, \hat{\tau}^{(b)}, \hat{\alpha}_1^{(b)}$, and $\hat{\alpha}_2^{(b)}$, update β and γ by the following equations:

$$\hat{\gamma}^{(b+1)} = \hat{\gamma}^{(b)} + \left\{ \sum_{i=1}^N \frac{\partial p_i^T}{\partial \gamma} \{V_{\gamma i}\}^{-1} \frac{\partial p_i}{\partial \gamma^T} \right\}^{-1} \sum_{i=1}^N \frac{\partial p_i^T}{\partial \gamma} \{V_{\gamma i}\}^{-1} (\hat{u}_i^{(b)} - p_i) \Big|_{(\gamma, \alpha_1) = (\hat{\gamma}^{(b)}, \hat{\alpha}_1^{(b)})}$$

$$\hat{\beta}^{(b+1)} = \hat{\beta}^{(b)} + \left\{ \sum_{i=1}^N \frac{\partial \lambda_i^T}{\partial \beta} \{V_{\beta i}\}^{-1} \text{Diag}(1 - u^{(b)}) \frac{\partial \lambda_i}{\partial \beta^T} \right\}^{-1} S_{\beta}$$

where $S_{\beta} = \sum_{i=1}^N \frac{\partial \lambda_i^T}{\partial \beta} \{V_{\beta i}\}^{-1} \text{Diag}(1 - u^{(b)}) (y_i - \lambda_i)$ and (β, τ, α_2) is replaced by $(\hat{\beta}^{(b)}, \hat{\tau}^{(b)}, \hat{\alpha}_2^{(b)})$.

Step 4: Given $\hat{\beta}^{(b+1)}$, $\hat{\gamma}^{(b+1)}$, $\hat{\tau}^{(b)}$, $\hat{\alpha}_1^{(b)}$ and $\hat{\alpha}_2^{(b)}$, update $\hat{\tau}$ by the following equation

$$\hat{\tau}^{(b+1)} = \frac{\sum_{i=1}^N \sum_{j=1}^{n_i} \lambda_{ij}^2 (1 - u_{ij})^2 ((y_{ij} - \lambda_{ij})^2 - \lambda_{ij})}{\sum_{i=1}^N \sum_{j=1}^{n_i} (1 - u_{ij})^2 \lambda_{ij}^4} \Big|_{(\beta, \gamma, \tau, u_{ij}) = (\hat{\beta}^{(b+1)}, \hat{\gamma}^{(b+1)}, \hat{\tau}^{(b)}, \hat{u}_{ij}^{(b)})}$$

Step 5: Given $\hat{\beta}^{(b+1)}$, $\hat{\gamma}^{(b+1)}$ and $\hat{\tau}^{(b+1)}$, α_1 can be updated by the following equation:

$$\hat{\alpha}_1^{(b+1)} = \hat{\alpha}_1^{(b)} + \left(\sum_{i=1}^n E_{\gamma i}^T W_{\gamma i}^{-1} E_{\gamma i} \right)^{-1} \left(\sum_{i=1}^n E_{\gamma i}^T W_{\gamma i}^{-1} (U_{\gamma i} - \rho_{\gamma i}(\alpha_1)) \right) \Big|_{(\gamma, u_{ij}) = (\hat{\gamma}^{(b+1)}, \hat{u}_{ij}^{(b)})}$$

In case that $W_{\gamma i}$ is the identity matrix, and $R_{\gamma i}(\alpha_1)$ has the compound symmetric structure, one could obtain the estimate α_1 as follows:

$$\hat{\alpha}_1 = \frac{\frac{1}{N^*} \sum_{i=1}^N \sum_{s < t} \frac{(u_{is} - p_{is})(u_{it} - p_{it})}{\sqrt{p_{is}(1 - p_{is})p_{it}(1 - p_{it})}}}{\frac{1}{N_{tot}} \sum_{i=1}^N \sum_{j=1}^{n_i} \frac{(u_{ij} - p_{ij})^2}{p_{ij}(1 - p_{ij})}} \Big|_{(\gamma, u_{ij}) = (\hat{\gamma}^{(b+1)}, \hat{u}_{ij}^{(b)})}$$

Step 6: Update α_2 by the following equation:

$$\hat{\alpha}_2^{(b+1)} = \hat{\alpha}_2^{(b)} + \left(\sum_{i=1}^N E_{\beta i}^T W_{\beta i}^{-1} H_{\beta i} E_{\beta i} \right)^{-1} S_{\beta}$$

where $S_{\beta} = \sum_{i=1}^N E_{\beta i}^T W_{\beta i}^{-1} H_{\beta i} (U_{\beta i} - \rho_{\beta i}(\alpha_2))$, and $(\beta, \gamma, \tau, u_{ij})$ in the right hand side expression is replaced by $(\hat{\beta}^{(b)}, \hat{\gamma}^{(b)}, \hat{\tau}^{(b)}, \hat{u}_{ij}^{(b)})$.

Step 7: Repeat Step 2 to 6 until $\hat{\beta}$ and $\hat{\gamma}$ converge.

Liang and Zeger (1986) proved that the GEE estimators of (β, γ) are consistent and asymptotic normal for any choice of working correlation matrix, provided that the regression model for the mean response has been correctly specified. The asymptotic covariance for the GEE estimator of (β, γ) is given by the sandwich form $B^{-1} M B^{-1}$

$$\text{Here } B = \begin{pmatrix} \sum_{i=1}^N \frac{\partial \hat{p}_i^T}{\partial \gamma} \{\hat{V}_{\gamma i}\}^{-1} \left(\frac{\partial \hat{p}_i}{\partial \gamma^T} \right) & 0 \\ 0 & \sum_{i=1}^N \frac{\partial \hat{\lambda}_i^T}{\partial \beta} \text{Diag}(1 - \hat{u}_i)^{1/2} \{\hat{V}_{\beta i}\}^{-1} \text{Diag}(1 - \hat{u}_i) \frac{\partial \hat{\lambda}_i}{\partial \beta^T} \end{pmatrix}$$

and

M =

$$\begin{pmatrix} \sum_{i=1}^N \frac{\partial \hat{p}_i^T}{\partial \gamma} \{\hat{V}_{\gamma i}\}^{-1} (\hat{u}_i - \hat{p}_i) \\ \sum_{i=1}^N \frac{\partial \hat{\lambda}_i^T}{\partial \beta} \{\hat{V}_{\beta i}\}^{-1} \text{Diag}(1 - \hat{u}_i) (\hat{y}_i - \hat{\lambda}_i) \end{pmatrix} \begin{pmatrix} \sum_{i=1}^N \frac{\partial \hat{p}_i^T}{\partial \gamma} \{\hat{V}_{\gamma i}\}^{-1} (\hat{u}_i - \hat{p}_i) \\ \sum_{i=1}^N \frac{\partial \hat{\lambda}_i^T}{\partial \beta} \{\hat{V}_{\beta i}\}^{-1} \text{Diag}(1 - \hat{u}_i) (\hat{y}_i - \hat{\lambda}_i) \end{pmatrix}^T$$

$$= \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}.$$

Here $M_{11} = \sum_{i=1}^N \frac{\partial \hat{p}_i^T}{\partial \hat{\gamma}} \{\hat{V}_{\gamma i}\}^{-1} (\hat{u}_i - \hat{p}_i) (\hat{u}_i - \hat{p}_i)^T \{\hat{V}_{\gamma i}\}^{-1} \frac{\partial \hat{p}_i}{\partial \hat{\gamma}^T}$,

$$M_{12} = M_{21}^T = \sum_{i=1}^N \frac{\partial \hat{p}_i^T}{\partial \hat{\gamma}} \{\hat{V}_{\gamma i}\}^{-1} (\hat{u}_i - \hat{p}_i) (\hat{y}_i - \hat{\lambda}_i)^T \text{Diag}(1 - \hat{u}_i) \{\hat{V}_{\beta i}\}^{-1} \frac{\partial \hat{\lambda}_i}{\partial \hat{\beta}^T},$$

$$M_{22} = \sum_{i=1}^N \frac{\partial \hat{\lambda}_i^T}{\partial \hat{\beta}} \{\hat{V}_{\beta i}\}^{-1} \text{Diag}(1 - \hat{u}_i) (\hat{y}_i - \hat{\lambda}_i) (\hat{y}_i - \hat{\lambda}_i)^T \text{Diag}(1 - \hat{u}_i) \{\hat{V}_{\beta i}\}^{-1} \frac{\partial \hat{\lambda}_i}{\partial \hat{\beta}^T}.$$

In the above expression, all quantities are replaced by the estimates at the convergence .

This variance estimate $B^{-1}MB^{-1}$ is often called robust estimate, or sandwich estimate. It should be noted that the asymptotic property holds when the number of subjects is larger.

When the number of subjects is small, correctly modeling the covariance matrix of Y_i could improve the efficiency of estimate. One issue with these formulas is that the true u_i are unknown and estimates are used. As a result the variance formula may not be accurate. We will revisit this issue at the simulation section.

The above algorithm has been implemented in R, which is shown in the Appendix.

The proposed method has been applied to analyze the dental caries data in Chapter 3, simulations to examine the performance of this method and the other two methods are carried out in Chapter 4.

CHAPTER III

DATA ANALYSIS

3.1 The three possible estimation methods for Iowa Fluoride Study data set

The Iowa fluoride study is designed to examine the effects of fluoride exposure in subsequent development of dental fluorosis and dental caries (Iowa Fluoride Study , 2013). The count of dental caries for each tooth ranged between 0 and 10. In the current project, we focused on a subset of Iowa Fluoride Study for children at the age of 5 years old. The missing observations (due to missing covariates) are assumed to be missing completely at random, and thus ignored. Since the dental caries counts for different teeth within a subject are potentially correlated, we applied a GEE model to analyze whether the dental caries is associated with fluoride exposure, adjusted by other covariates, such as gender, age, daily soda pop intake etc. as described in the section 1.2.

The GEE method has been implemented as a function *geeglm* in R-package “geepack”. The function *geeglm* has been used to analyze the dental data, where the compound symmetric correlation structure (i.e. $\text{corr}(Y_{ij}, Y_{ij'}) = \alpha$, a constant for $j \neq j'$) has been used. The estimated parameters (*Est.*), their standard error (*Std. Err.*), and p-value based on the Wald test are reported in Table 1 under the column title “GEE Poisson”. A GEE does not model excessive zeros. To account for excessive zeros, a zero-inflated NB (ZINB) has been applied to analyze the Iowa Fluoride data, and the results

are reported in Table 1 and Table 2 under the column title “ZINB”. The function used was *zeroinf* under the R-package “pscl”. In the zero-inflated NB model, the correlations for observations from the same subject are ignored. The GEE based zero-inflated NB (GEE.ZINB) proposed is more general than both zero inflated NB model and the GEE model since both the excessive zeros and the correlations of observations from the same subject are considered. The R code for the algorithm is reported in Appendix B, and the results for analyzing the fluoride study are shown under the column title “GEE.ZINB” in Table 1 and Table 2.

3.2 Analysis results for Iowa Fluoride Study data set

The data set analyzed in this project is a subset of Iowa Fluoride Study of children at the age of 5 years old. After removing the observations with missing covariates, the study subset includes 414 subjects with 8189 observations in total. The results based on the three different methods are summarized in Table 1 for models including all the covariates. Based on Table 1, the magnitude and the direction (i.e. the sign of the estimates) based on the three different methods are similar. The correlation coefficient for the observations within the same subject are 0.166 for count of caries in the GEE based ZINB model and 0.275 for zero-inflated component in the same model, and the correlation coefficients for the within-subject observations based on GEE model is 0.129. The small correlation coefficients in all models indicate that the correlations for within subject observations are not high, thus, the result from ZINB which ignore the correlations may not be as serious as those from GEE which ignore excessive zeros counts. Thus, the results based on GEE.ZINB could be similar to those based on ZINB.

Indeed, the result based on ZINB model and the results based on GEE.ZINB are very close (see Table 1). All three models indicate that the “DentalExamAge” has significant impact on numbers of caries, the larger of the “DentalExamAge”, the larger numbers of caries. Both ZINB and GEE.ZINB models indicate that (i) “AUCmgFO_5yrs” is negatively associated with number of carries (see estimates under count model); (ii) “ToothBrushingFreqPerDayAvg” is positively associated with the probability of zero-count, the more frequent to brush teeth, the higher probability of zero counts, indicating less possibility to have caries. Based on GEE.ZINB, we also found that the “DentalVistPast6moAvg” is positively associated with number of caries ($P.value=0.015$), and “FluorideTreatment6moAvg” is also positively associated with number of caries at the significant level 0.069.

Based on the results from GEE.ZINB shown in Table 1, four covariates (i.e., “DentalExamAge”, “AUCmgFO_5yrs”, “DentalVisitPast6moAvg”, “FluorideTreatment6moAvg”) are significantly associated with NB count component, and only one covariate (i.e. “ToothBrushingFreqPerDayAvg”) related to zero-inflation. We fit reduced models with these significant covariates in the each of three models (GEE.ZINB, ZINB, GEE) and the results are shown in Table 2. Based on the results from Table 2, all covariates are significant in the GEE.ZINB model. In addition, the ZINB and GEE models found that (i) “DentalExamAge” is significantly positively associated with count of caries, and (ii) “AUCmgFO_5yrs” is significantly negatively associated with count of caries. In addition, both GEE.ZINB and ZINB found that “FluorideTreatment6moAvg” is significantly positively associated with count of caries.

Based on Table 1 and Table 2, we also found that the standard errors resulted from GEE is larger than those obtained from GEE.ZINB and ZINB models for almost all parameters in the count models, indicating that the inferences based on GEE model are less powerful if the excessive zeros do exist. We also found that the differences for the estimated parameters based on GEE.ZINB and GEE are relatively larger. To examine the performances of GEE.ZINB, ZINB and GEE, simulation studies are carried out in the next chapter.

Table 1: The analysis results for Iowa Fluoride Study data set based on three different models: the GEE based zero-inflated NB model (see the result under the title “GEE.ZINB”), the zero-inflated NB model (see the result under the title “ZINB”), and the GEE Poisson model (see the result under the title “GEE Poisson”).

	GEE.ZINB		ZINB		GEE Poisson
	Count Model	Zero-Inflated	Count Model	Zero-Inflated	Count Model
Est.(Intercept)	-3.5062	2.897	-3.6322	2.7143	-4.6990
Std. Error	0.903	1.028	0.886	0.778	1.4992
P-Value	(<0.001)	(0.0048)	(<0.001)	(<0.001)	(0.0017)
Est.(GenderM1)	-0.0108	0.207	-0.0044	0.2099	-0.1616
Std. Error	0.133	0.130	0.149	0.128	0.2574
P-Value	(0.936)	(0.1115)	(0.976)	(0.01)	(0.5303)
Est.(DentalExamAge)	0.806	-0.306	0.8209	-0.2869	0.7584
Std. Error	0.170	0.192	0.16	0.137	0.2824
P-Value	(<0.005)	(0.1106)	(<0.001)	(0.037)	(0.0072)
Est.(AUCmgF0_5yrs)	-0.936	0.351	-0.9063	0.4136	-1.1949
Std. Error	0.247	0.324	0.308	0.294	0.6187
P-Value	(<0.001)	(0.2792)	(0.003)	(0.159)	(0.0534)
Est.(AUCSodaOz0_5yrs)	0.0808	-0.005	0.0783	-0.0084	0.0375
Std. Error	0.053	0.036	0.06	0.046	0.0808
P-Value	(0.125)	(0.9018)	(0.194)	(0.856)	(0.6426)
Est.(ToothBrushingFreqPerDayAvg)	-0.0467	0.513	-0.0576	0.5246	-0.4447
Std. Error	0.168	0.167	0.201	0.172	0.3108
P-Value	(0.782)	(0.0021)	(0.774)	(0.002)	(0.1525)
Est.(DentalVisitPast6moAvg)	0.6936	0.903	0.8335	0.8911	0.0506
Std. Error	0.285	0.563	0.525	0.513	1.0703
P-Value	(0.015)	(0.1090)	(0.112)	(0.082)	(0.9623)
Est.(FluorideTreatment6moAvg)	1.222	-0.717	1.234	-0.6758	1.6111
Std. Error	0.672	0.703	0.727	0.63	1.2019
P-Value	(0.069)	(0.3083)	(0.09)	(0.283)	(0.1801)
Est.(HomeFluorideppmAvg)	-0.0701	0.032	-0.0883	0.0111	-0.1906
Std. Error	0.132	0.126	0.176	0.135	0.2540
P-Value	(0.595)	(0.7980)	(0.615)	(0.935)	(0.4531)
α	0.166	0.275	N/A	N/A	0.129
τ	0.455	N/A	0.928	N/A	N/A

Table 2: The analysis results of the reduce model based on three different models: the GEE based zero-inflated NB model (see the result under the title “GEE.ZINB”), the zero-inflated NB model (see the result under the title “ZINB”), and the GEE Poisson model (see the result under the title “GEE Poisson”).

	GEE.ZINB		ZINB		GEE Poisson
	Count Model	Zero-Inflated	Count Model	Zero-Inflated	Count Model
Est.(Intercept)	-3.923	1.783	-4.586	1.506	-5.454
Std. Error	0.243	0.170	0.832	0.213	1.553
P-Value	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
Est.(DentalExamAge)	0.935	N/A	1.029	N/A	0.823
Std. Error	0.043	N/A	0.154	N/A	0.283
P-Value	(<0.001)	N/A	(<0.001)	N/A	(0.003)
Est.(AUCmgF0 5yrs)	-1.163	N/A	-1.241	N/A	-1.499
Std. Error	0.074	N/A	0.228	N/A	0.607
P-Value	(<0.001)	N/A	(<0.001)	N/A	(0.013)
Est.(DentalVisitPast6moAvg)	0.565	N/A	0.505	N/A	-0.009
Std. Error	0.175	N/A	0.468	N/A	1.060
P-Value	0.0012	N/A	(0.280)	N/A	(0.992)
Est.(FluorideTreatment6moAvg)	1.278	N/A	1.424	N/A	1.592
Std. Error	0.201	N/A	0.640	N/A	1.178
P-Value	(<0.001)	N/A	(0.026)	N/A	(0.176)
Est.(ToothBrushingFreqPerDayAvg)	N/A	0.551	N/A	0.593	N/A
Std. Error	N/A	0.166	N/A	0.142	N/A
P-Value	N/A	(<0.001)	N/A	(<0.001)	N/A
α	0.179	0.255	N/A	N/A	0.127
τ	0.341	N/A	0.732	N/A	N/A

CHAPTER IV

A SIMULATION STUDY

We have proposed the GEE-based zero-inflated NB (GEE.ZINB) model to incorporate both the correlations of within-subject observations, and the excessive zero counts. It is noticed that GEE incorporate is the correlations of within-subject observations but ignores the excessive zero counts, while the zero- inflated NB model (ZINB) incorporate is the excessive zero counts but ignores the possible correlations of the within-subject observations. To examine the performance is of the three methods, we constructed the following three scenarios. The underlying true model was the GEE based zero-inflated NB model obtained from the reduce model in the previous chapter. That is, we assume that the count model, i.e., the number of caries for a tooth, follows a log-linear model:

$$\begin{aligned} \log(\lambda_{ij}) = & 3.923 + 0.935DentalExamAge - 1.163AUCmgF0_5yrs \\ & + 0.565DentalVisitPast6moAvg + 1.278FluorideTreatment6moAvg, \end{aligned} \quad (23)$$

and the probability of zero-count from the zero-degenerated distribution follows the following logistic regression model:

$$\text{logit } P_{ij} = 1.783 + 0.551ToothBrushingFrequencyPerDayAvg. \quad (24)$$

We used the covariates in the first 200 subjects in the dental data set, and generated 20 correlated binary random variables for each subject, where the binary random variables were the quantiles of the Bernoulli distribution with mean specified in the logit model (24) for the subject and the probabilities as those specified by 20 correlated standard normal random numbers with an intra-class correlation coefficient α . Meanwhile, we also generated 20 NB random variables for each subject, where the 20 NB random numbers were the quantiles of the NB distribution ($\tau=0.341$) with mean specified in the log-linear model (23) and the probabilities as those specified by 20 correlated standard normal random numbers with the intra-class correlation coefficient α . For each subject, the number of caries for each tooth is set as zero if the associated binary variable is 1, indicating that the observation is from zero-degenerated distribution; the number of caries is set as the count from the associated NB random number, if the binary variable is zero, indicating the observation is from NB distribution. The total number of observations in each simulated data set is 4000 observed in 200 subjects. By setting α at 0.9, 0.5, and 0.0, respectively, we simulated the following three scenarios: excessive zeroes with high correlated counts for caries (Scenario 1), excessive zeroes with medium correlated counts for caries (Scenario 2), and excessive zeroes with uncorrelated counts for caries (Scenario 3). The simulation results are reported in Table 3, 4, 5 for the three different scenarios, respectively. The result for each scenario was obtained from 1000 generated data sets; Under each scenario, we fitted each data set with the three different models: GEE based zero-inflated NB model (GEE.ZINB), zero-inflated NB model (ZINB), and the GEE quasi-Poisson model. The simulation results were summarized by the means of the estimated parameters (MEAN), the average of the estimated sandwich variance (VAR.M),

the variance of the 1000 estimates for each parameter (VAR.S), and the predicted mean squared errors (PMSE). The PMSE is defined as the average of the squared differences of the estimates and the underlying values.

Based on the simulation results, we conclude that (i) the predicted mean squared error (PMSE) are reduced for all three methods when the correlations changed from high (0.9, Table 3) to moderate (0.5, Table 4) and to uncorrelated (0.0, Table 5) scenarios, and the GEE.ZINB has the smallest PMSE among the three models under each scenario; (ii) the GEE.ZINB are stable in estimating both β and γ , while ZINB is unstable in estimating γ , which is indicated by the larger variance in the simulated γ , particularly when the correlation coefficient is high (see Table 3). From the simulation results, we conclude that the GEE.ZINB performed better than ZINB and GEE under all the three scenarios in terms of having smaller PMSE and are stable in estimating all parameters. However, we also noticed that there is a large discrepancy between the sandwich estimates (VAR.M) and the simulated sample variances (VAR.S), which may be caused by ignoring the variation introduced by the latent variable indicator variable. A formula to consider the variation caused by estimating the latent variable may be a remedy to the problem. As an alternative, the nonparametric bootstrap methods could be applied to obtain the variance estimates. In next chapter, we reanalyzed the dental caries data set, where the variances were estimated from nonparametric bootstrap method.

Table 3: Simulation results for scenario 1, where excessive zeroes exist and the counts are highly correlated.

Coefficients	True	GEE.ZINB				ZINB				GEE			
Count Model		MEAN	VAR.M	VAR.S	PMSE	MEAN	VAR.M	VAR.S	PMSE	MEAN	VAR.M	VAR.S	PMSE
(Intercept)	-3.923	-4.120 (-4.315)*	0.288 (0.128)*	13.701	14.774 (6.240)*	-4.119 (-4.043)*	1.215 (0.999)*	15.67	15.696 (5.376)*	-4.677 (-4.400)*	13.110 (5.262)*	12.553	13.110 (5.262)*
DentalExam Age	0.935	0.921 (0.958)*	0.010 (0.003)*	0.481	0.520 (0.173)*	0.899 (0.925)*	0.042 (0.034)*	0.517	0.518 (0.159)*	0.626 (0.595)*	0.529 (0.270)*	0.434	0.529 (0.270)*
AUCmgF0_ 5yrs	-1.163	-1.389 (-1.359)*	0.038 (0.019)*	1.526	1.554 (0.634)*	-1.358 (-1.321)*	0.119 (0.107)*	1.581	1.617 (0.641)*	-1.411 (-1.259)*	1.624 (0.665)*	1.565	1.624 (0.665)*
DentalVisitP ast6moAvg	0.565	0.447 (0.382)*	0.111 (0.058)*	4.853	5.041 (2.074)*	0.398 (0.426)*	0.380 (0.338)*	5.025	5.048 (1.876)*	0.341 (0.466)*	5.701 (2.443)*	5.675	5.701 (2.443)*
FluorideTre atment6mo Avg	1.278	1.264 (1.301)*	0.170 (0.072)*	7.236	7.788 (2.886)*	1.195 (1.165)*	0.549 (0.467)*	7.600	7.599 (3.094)*	0.984 (0.965)*	9.108 (4.031)*	9.031	9.108 (4.031)*
Zero-Inflated Model													
(Intercept)	1.783	1.343 (1.347)*	0.086 (0.077)*	0.472	0.679 (0.335)*	-2.669 (1.626)*	42746 (0.041)*	4765	4780 (0.350)*	N/A	N/A	N/A	N/A
ToothBrushi ngFreqPerD ayAvg	0.551	0.637 (0.598)*	0.081 (0.069)*	0.477	0.490 (0.216)*	-2.945 (0.501)*	295197 (0.038)*	4187	4196 (0.343)*	N/A	N/A	N/A	N/A

Note: The number inside the parenthesis with * indicates median instead of mean. VAR.M is the variance of model. VAR.S is the variance of sample.

Table 4: Simulation results for scenario 2, where excessive zeroes exist and the counts are moderately correlated.

Coefficients	True	GEE.ZINB				Zero-Inflated NB				GEE			
Count Model		MEAN	VAR.M	VAR.S	PMSE	MEAN	VAR.M	VAR.S	PMSE	MEAN	VAR.M	VAR.S	PMSE
(Intercept)	-3.923	-4.255 (-4.275)*	0.090 (0.059)*	3.096	3.206 (1.12)*	-4.334 (-4.094)*	0.867 (0.794)*	3.978	4.147 (1.185)*	-5.555 (-5.396)*	2.212 (2.045)*	6.350	9.013 (3.731)*
DentalExama ge	0.935	0.968 (0.984)*	0.002 (0.001)*	0.106	0.107 (0.04)*	0.939 (0.934)*	0.029 (0.026)*	0.100	0.100 (0.032)*	0.801 (0.759)*	0.065 (0.061)*	0.226	0.244 (0.130)*
AUCmgF0_5y rs	-1.163	-1.249 (-1.254)*	0.017 (0.014)*	0.334	0.341 (0.169)*	-1.254 (-1.231)*	0.099 (0.094)*	0.338	0.346 (0.155)*	-1.346 (-1.283)*	0.465 (0.391)*	0.554	0.587 (0.238)*
DentalVisitPas t6moAvg	0.565	0.512 (0.546)*	0.047 (0.038)*	0.142	0.144 (0.505)*	0.515 (0.521)*	0.324 (0.307)*	1.086	1.088 (0.461)*	0.514 (0.531)*	1.612 (1.446)*	1.864	1.866 (0.875)*
FluorideTreat ment6moAvg	1.278	1.342 (1.337)*	0.060 (0.047)*	1.818	1.822 (0.683)*	1.281 (1.258)*	0.467 (0.433)*	1.741	1.741 (0.652)*	1.212 (1.188)*	2.394 (2.180)*	3.213	3.217 (1.323)*
Zero-Inflated Model													
(Intercept)	1.783	1.553 (1.559)*	0.042 (0.039)*	0.164	0.217 (0.094)*	0.562 (1.723)*	69575 (0.040)*	1610.5	1612 (0.124)*	N/A	N/A	N/A	N/A
ToothBrushin gFreqPerDay Avg	0.551	0.565 (0.550)*	0.039 (0.035)*	0.163	0.163 (0.076)*	-4.462 (0.460)*	297048 (0.035)*	1961.9	1987 (0.119)*	N/A	N/A	N/A	N/A

Note: The number inside the parenthesis with * indicates median instead of mean. VAR.M is the variance of model. VAR.S is the variance of sample.

Table 5: Simulation results for scenario 3, where excessive zeroes exist and the counts are uncorrelated.

Coefficients	True	GEE.ZINB				ZINB				GEE			
Count Model		MEAN	VAR.M	VAR.S	PMSE	MEAN	VAR.M	VAR.S	PMSE	MEAN	VAR.M	VAR.S	PMSE
(Intercept)	-3.923	-3.855 (-3.845)*	19.984 (0.287)*	0.522	0.527 (0.216)*	-4.365 (-4.066)*	0.713 (0.597)*	1.699	1.894 (0.459)*	-6.186 (-6.299)*	0.503 (0.333)*	1.930	7.051 (5.649)*
DentalExam Age	0.935	0.927 (0.932)*	0.806 (0.009)*	0.016	0.016 (0.006)*	0.930 (0.929)*	0.023 (0.019)*	0.020	0.020 (0.008)*	0.930 (0.960)*	0.014 (0.007)*	0.070	0.070 (0.034)*
AUCmgF0_ 5yrs	-1.163	-1.178 (-1.177)*	0.115 (0.049)*	0.088	0.089 (0.037)*	-1.219 (-1.215)*	0.093 (0.087)*	0.106	0.109 (0.046)*	-1.345 (-1.332)*	0.113 (0.108)*	0.122	0.155 (0.060)*
DentalVisit Past6moAv g	0.565	0.529 (0.517)*	0.330 (0.162)*	0.265	0.266 (0.122)*	0.539 (0.529)*	0.308 (0.278)*	0.291	0.292 (0.142)*	0.569 (0.567)*	0.396 (0.379)*	0.394	0.394 (0.177)*
FluorideTre atment6mo Avg	1.278	1.271 (1.294)*	2.646 (0.242)*	0.370	0.370 (0.169)*	1.261 (1.288)*	0.435 (0.380)*	0.418	0.418 (0.188)*	1.306 (1.305)*	0.569 (0.537)*	0.632	0.633 (0.307)*
Zero-Inflated Model													
(Intercept)	1.783	1.769 (1.763)*	0.013 (0.013)*	0.033	0.033 (0.016)*	-0.781 (1.735)*	452665. 2 (0.039)*	24.486	31.060 (0.031)*	N/A	N/A	N/A	N/A
ToothBrushi ngFreqPerD ayAvg	0.551	0.553 (0.556)*	0.013 (0.012)*	0.031	0.031 (0.013)*	-1.053 (0.480)*	197473 7.2 (0.034)*	10.570	13.143 (0.027)*	N/A	N/A	N/A	N/A

Note: The number inside the parenthesis with * indicates median instead of mean. VAR.M is the variance of model. VAR.S is the variance of sample.

CHAPTER V

A REANALYSIS OF THE DENTAL DATA USING THE BOOTSTRAP METHOD

From the simulation results presented in the previous chapter, the sandwich variance estimates are smaller than the simulated sample variances in all the three scenarios. Therefore, the sandwich variance estimate presented in Chapter II underestimate is the true variance. To make correct inferences, the variance estimates should be corrected. One may re-examine the theoretical derivation for the sandwich estimate. An alternative method is using the bootstrap technique to estimate the variance. To do the bootstrap sampling, we draw a random sample (with replacement, subject as sampling unit) from the observed data set, and fit the data with the three models. This process is repeated for 1000 times. For each parameter, the sample variance for the 1000 estimates is the variance of the estimated parameter. We reanalyzed the dental caries data, and the results are reported in Table 6. Based on the Table 6, “DentalExamAge” and “AUCmgF0_5yrs” are significantly associated with number of caries.

Table 6: The analysis results for Iowa Fluoride Study based on bootstrapped variance and three different models: the GEE based zero-inflated NB model (“GEE.ZINB”), the zero-inflated NB model (“ZINB”), and the GEE Poisson model (“GEE Poisson”).

	GEE.ZINB		ZINB		GEE Poisson
	Count Model	Zero-Inflated	Count Model	Zero-Inflated	Count Model
Est.(Intercept)	-3.9035	3.0844	-4.7266	-1.457	-5.2236
(Boot strap based) Std. Error	1.5524	1.8997	2.293	34.88	1.9743
P-Value	(0.0232)	(0.124)	(0.1132)	(0.938)	(0.0173)
Est.(GenderM1)	0.0241	0.2214	0.0546	0.136	-0.1752
(Boot strap based) Std. Error	0.2730	0.2761	0.359	3.39	0.2747
P-Value	(0.9617)	(0.455)	(0.9902)	(0.951)	(0.5564)
Est.(DentalExamAge)	0.8606	-0.3556	0.9744	0.456	0.8553
(Boot strap based)Std. Error	0.2870	0.3451	0.389	7.81	0.3780
P-Value	(0.0048)	(0.371)	(0.0348)	(0.971)	(0.0448)
Est.(AUCmgF0_5yrs)	-0.9399	0.4590	-1.1051	-0.926	-1.2714
(Boot strap based)Std. Error	0.4716	0.7182	0.709	16.04	0.6669
P-Value	(0.0408)	(0.675)	(0.2014)	(0.979)	(0.0732)
Est.(AUCSodaOz0_5yrs)	0.0917	-0.0164	0.0722	-0.356	0.0642
(Boot strap based)Std. Error	0.0841	0.0793	0.152	3.38	0.0954
P-Value	(0.3184)	(0.991)	(0.6053)	(0.998)	(0.6943)
Est.(ToothBrushingFreqPerDayAvg)	-0.0134	0.5478	0.0061	1.032	-0.4849
(Boot strap based)Std. Error	0.3303	0.3540	0.450	4.25	0.3344
P-Value	(0.8976)	(0.154)	(0.8981)	(0.902)	(0.1836)
Est.(DentalVisitPast6moAvg)	0.9685	0.9324	1.0570	0.206	-0.0173
(Boot strap based)Std. Error	0.7800	1.0035	0.991	25.83	1.1203
P-Value	(0.2952)	(0.357)	(0.4001)	(0.972)	(0.9640)
Est.(FluorideTreatment6moAvg)	1.1871	-0.6378	1.2206	-1.496	1.5956
(Boot strap based)Std. Error	1.3809	1.3192	1.682	14.50	1.2104
P-Value	(0.3684)	(0.585)	(0.4632)	(0.963)	(0.1832)
Est.(HomeFluorideppmAvg)	-0.0555	-0.0034	0.0105	0.372	-0.1667
(Boot strap based)Std. Error	0.3514	0.3083	0.542	4.39	0.3015
P-Value	(0.8658)	(0.883)	(0.8705)	(0.998)	(0.5247)
α	0.166	0.275	N/A	N/A	0.129
τ	0.455	N/A	0.928	N/A	N/A

CHAPTER VI

DISCUSSION AND FUTURE WORK

We have developed GEE based zero-inflated models for the dental caries data, where excessive zeroes are apparent and the numbers of caries from the teeth within a subject are correlated. Simulations are carried out to illustrate that this proposed method performed better than the independent zero-inflated model and the standard GEE model.

Although the test statistics for the regression parameters or the functions of the parameters could be easily constructed, their performances need to be further examined through a power study. In addition, the derivation of the sandwich variance formula needs to be examined so that a proper formula which incorporates the variation in estimating the latent variable could be developed. Furthermore, we plan to extend the GEE based zero-inflated negative binominal model to a zero inflated Conell-Maxwell-Power (CMP) model (Kimberly, Borle, & Shmueli, 2012) with the hope that this model may fit the data even more accurately.

REFERENCES

- Iowa Fluoride Study* . (2013, April 10). Retrieved from Iowa Fluoride Study :
[http://www.dentistry.uiowa.edu/preventive/research/fluoride study.shtml](http://www.dentistry.uiowa.edu/preventive/research/fluoride%20study.shtml)
- Albert, P., Zeger, S. L., & Liang, K. Y. (1988). Models for longitudinal data: a generalized estimating. *Biometrics*, 44, 1049–1060.
- Broffitt, B., Levy , S. M., Warren, J. J., & Cavanaugh, J. E. (2007). An investigation of bottled water use and caries in the mixed dentition. *Journal of Public Health Dentistry*, 67, 151-158.
- Carey, V. J., Lumley, T., & Ripley, B. (1998, January 27). Generalized Estimation Equation solver. (4.13-18). Retrieved from <http://cran.r-project.org/web/packages/gee>
- Chankanka, O. (2010). Dietary intake and dental caries in children. *PhD dissertation, University of Iowa*.
- Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied Longitudinal Analysis* (2nd ed.). New York: Wiley.
- Hall, D. B., & Zhang, Z. (2004). Marginal models for zero inflated clustered data. *Statistical Modelling*, 4, 161-180.
- Hedeker, D. (2004). An Introduction To Growth Modeling. In D. Kaplan, & M. Seltzer. *Handbook of Quantitative Methodology for the Social Sciences* (pp. 215-234). Thousand Oak CA: Sage Publications.
- Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal Data Analysis*. New Jersey, NY: Wiley & Sons, Inc.
- Hong, L., Levy , S. M., Warren, J. J., & Broffitt, B. (2009). Association between Enamel Hypoplasia and Dental Caries in Primary Second Molars: A Cohort Study. *Carries Research*, 43, 345-353.
- Iida, H., & Kumar, J. V. (2009). The association between enamel fluorosis and dental caries in U.S. schoolchildren. *Journal of the American Dental Association*, 140, 855-862.
- Jackman, S., Tahk, A., Zeileis, A., Maimone, C., & Fearon, J. (2012, June 13). Political Science Computational Laboratory, Stanford University. (1.04.4). Retrieved from <http://pscl.stanford.edu/>

- Jun, Y., & Halekoh, U. (2012, January 27). Generalized Estimating Equation Package. (1.1-6). Retrieved from <http://cran.r-project.org/web/packages/geepack>
- Kimberly, F. S., Borle, S., & Shmueli, G. (2012). The COM-Poisson Model for count data: A survey of methods and applications. *Applied Stochastic Models in Business and Industry*, 28, 128-129.
- Levy, S. M., Warren, J. J., Broffitt, B., Hillis, S. L., & Kanellis, M. J. (2003). Fluoride, beverages and dental caries in the primary dentition. *Caries Research*, 37, 157-165.
- Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.
- Marshall, T. A., Broffitt, B. A., Eichenberger-Gilmore, J., Warren, J. J., Cunningham, M. A., & Levy, S. M. (2005). The roles of meal, snack, and daily total food and beverage exposures on caries experience in young children. *Journal of Public Health Dentistry*, 65, 166-173.
- Marshall, T. A., Eichenberger-Gilmore, J. M., Larson, M. A., Warren, J. J., & Levy, S. M. (2007a). Comparison of the intakes of sugars by young children with and without dental caries experience. *Journal of the American Dental Association*, 138, 39-46.
- Marshall, T. A., Eichenberger-Gilmore, J. M., Broffitt, B. A., Warren, J. J., & Levy, S. M. (2007b). Dental caries and childhood obesity: roles of diet and socioeconomic status. *Community Dent Oral Epidemiol*, 35, 449-458.
- Mayo foundation for Medical Education and Research. (2012, August 10). Children's health. Retrieved from <http://www.mayoclinic.com/health/medical/IM03923>
- Prentice, R. (1988). Correlated binary regression with covariate specific to each binary/ observation. *Biometrics*, 44, 1033-1048.
- R Development Core Team. (2013). R: A Language and Environment for Statistical Computing. (Version 2.15.3). Vienna, Austria. Retrieved from <http://www.R-project.org>
- Wan, T., Hua, H., & Xin, T. (2012). *Applied Categorical and Count Data Analysis*. New York : Chapman and Hall/CRC.
- Wikipedia encyclopedia*. (n.d.). Retrieved from Wikipedia : http://en.wikipedia.org/wiki/Dental_caries#cite_note-medline-1
- Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression Models for Count Data in R. *Journal Of Statistical Software*, 27, 1-25.

APPENDIX A

R code for data adjustment and Figure 2 and Figure 3

```
##### Data Input #####

Caries <- read.csv("c:\\R\\IowaFluorideStudy_Age5_Caries.csv",header=T)
Demographics <- read.csv("c:\\R\\IowaFluorideStudy_Demographics.csv",
                        header=T)
SupportingData <-read.csv("c:\\R\\IowaFluorideStudy_Age5_SupportingData.csv",
                        header=T)

#####
##                               Step 1: Combinate to dental
#####

##### Fixed Table1.Demographics~~~~Dental.Demographics #####

attach(Demographics)
names(Demographics)

##### separate income2007 in two parts #####

NewIncome2007 <- strsplit(as.character(Demographics$Income2007), "[.]")
Income2007Char <- sapply(NewIncome2007,function(x) x[1])
Num <- strsplit(sapply(NewIncome2007,function(x) x[2]),"[$]")
Income2007Num <- sapply(Num ,function(x) x[2])

##### separate Race in two parts #####

for(i in 1:length(Demographics$Race))
{
  while(Demographics$Race[i]=="4.Hispanic")
  {
    Demographics$Race[i]="1.White"
  }
}
NewRace <- strsplit(as.character(Demographics$Race), "[.]")
RaceChar <- sapply(NewRace,function(x) x[2])
RaceNum <- sapply(NewRace,function(x) x[1])

##### Make the table Dental.Demographics #####
```

```

Dental.Demographics <-
cbind(Demographics[,1:2],Income2007Char,Income2007Num,RaceChar,RaceNum)

#### Fixed Table2.Caries~~~~Dental.Caries ####

head(Caries)

Caries <- data.frame(Caries)
CariesChange <- rep(NA,length(Caries$Caries))

#### Caries s,d0 equal '0', d1 equal '1', d2,f equal '2' ####

i <- grep("s",Caries$Caries)
j <- grep("d0",Caries$Caries)
k <- grep("d1",Caries$Caries)
m <- grep("d2",Caries$Caries)
n <- grep("f",Caries$Caries)
CariesChange[i] = 0
CariesChange[j] = 0
CariesChange[k] = 1
CariesChange[m] = 2
CariesChange[n] = 2

#### Product a new tabs Dental.Caries ####

Dental.Caries <- cbind(Caries[,1:3],CariesChange,Caries[,5:6])

#### Combinate tables 1,2,4-----dental ####

attach(SupportingData)
names(SupportingData)

m1 <- merge(Dental.Demographics,SupportingData,by.x='SID')
dental <- merge(m1,Dental.Caries,by.x='SID')

#####
## Step2: Count number of the caries
#####

attach(dental)
names(dental)

#### Check how many Teeth in the table ####

teamNo <- subset(dental, dental$Surface=="b")

```

```
##### To Count the Number of the caries for each tooth #####
```

```
CariesCount <- rep(NA, length(teamNo$Surface))
totalID <- unique(dental$SID)
j=1
for(ID in totalID)
{
  tempdata <- dental[dental$SID==ID,]
  teeth <- unique(tempdata$Tooth)
  for(i in teeth)
  {
    tempteeth <- tempdata[tempdata$Tooth==i,]
    CariesCount[j] <- sum(tempteeth$CariesChange, na.rm = T)
    j=j+1
  }
}
```

```
#####
## Step 3: To Get The New Table---NewDental
#####
```

```
m1 <- merge(Dental.Demographics, SupportingData, by.x='SID')
```

```
##### To Get the SID Tooth CariesCount For Each Tooth #####
```

```
SidTeeth <- unique(paste(dental$SID, dental$Tooth))
SplitSidTeeth <- strsplit(SidTeeth, "[ ]")
```

```
TotalTeeth <- length(teamNo$Surface)
SID <- rep(NA, TotalTeeth)
Tooth <- rep(NA, TotalTeeth)
for (i in 1:TotalTeeth)
{
  SID[i] <- SplitSidTeeth[[i]][1]
}
for (i in 1:TotalTeeth)
{
  Tooth[i] <- SplitSidTeeth[[i]][2]
}
```

```
##### To Get the NewDental Table #####
```

```
tempTable <- cbind(SID, Tooth, CariesCount)
NewDental <- merge(tempTable, m1, by.x='SID')
head(NewDental)
```

```

#####      Polt the Caries vs. teeth      #####

n0=0;n1=0;n2=0;n3=0;n4=0;n5=0;n6=0;n7=0;n8=0;n9=0;n10=0;

for(i in 1:TotalTeeth)
{
  if(CariesCount[i]==0) {n0=n0+1}; if(CariesCount[i]==5) {n5=n5+1}
  if(CariesCount[i]==1) {n1=n1+1}; if(CariesCount[i]==6) {n6=n6+1}
  if(CariesCount[i]==2) {n2=n2+1}; if(CariesCount[i]==7) {n7=n7+1}
  if(CariesCount[i]==3) {n3=n3+1}; if(CariesCount[i]==8) {n8=n8+1}
  if(CariesCount[i]==4) {n4=n4+1}; if(CariesCount[i]==9) {n9=n9+1}
  if(CariesCount[i]==10){n10=n10+1}
}

#####      Figure 2      #####

count <- matrix(c(n0,n1,n2,n3,n4,n5,n6,n7,n8,n9,n10),ncol=11)
colnames(count)<- c("0","1","2","3","4","5","6","7","8","9","10")
barplot(count, xlab="The Number of the caries ", ylab="The number of teeth")

#####      Figure 3      #####

count1 <- matrix(c(n1,n2,n3,n4,n5,n6,n7,n8,n9,n10),ncol=10)
colnames(count1)<- c("1","2","3","4","5","6","7","8","9","10")
barplot(count1, xlab = "The Number of the caries",ylab="The number of teeth")

```

APPENDIX B

R code for GEE.ZINB

```
#### Data input ####
```

```
Dental <- read.csv("F:\\R\\IowaF.csv", header=T)
head(Dental)
install.packages("pscl")
library(pscl)
install.packages("MASS")
library(MASS)
install.packages("gee")
library(gee)
install.packages("Matrix")
library(Matrix)
```

```
#####
##                                     GEE ZINB
#####
```

```
RowbyRow<-function(A, b)
{temp<-A
  for (i in 1:length(A[,1]))
    {temp[i,]<-A[i,]*b[i]}
  return(temp)
}
```

```
GEE.ZINB.Single.Iter <- function(X, Y, Z, beta0, gamma0, tao0=tao0,
                                alpha10=0.5, alpha20=0.5)
```

```
{
  IDS <- unique(X$SID)
  X1<-X[, c(-1)]; X1<-as.matrix(cbind(rep(1,length(X[,1])), X1))
  Z1<-Z[, c(-1)]; Z1<-as.matrix(cbind(rep(1,length(Z[,1])), Z1))
  ## attributes(X1)
  n.X <- length((X1[,1]))
  n.Z <- length((Z1[,1]))
  DX<-X1; DZ<-Z1
  Y1 <- as.vector(Y[, c(-1)])
  lamda1 <- exp(X1%*%beta0)
  eta1 <- exp(Z1%*%gamma0)
  p1<-eta1/(1+eta1)
```

```
####      Update Uij      ####
```



```

temp<-(1-p1)*(1+tao0*lamda1)^(-(1/tao0))
ub<-ifelse(Y1==0,1,0)/(1+temp/p1)

## sum(ub==0); sum(Y1!=0)

####      Update gamma and beta      ####

DevBeta <- lamda1
DevGamma <- p1*(1-p1)
DX<-RowbyRow(X1, DevBeta)
DZ<-RowbyRow(Z1, DevGamma)
V.gamma<-p1*(1-p1)
V.beta<-lamda1*(1+tao0*lamda1)
dvd.gamma<- matrix(rep(0,n.Z*n.Z), nrow=n.Z, ncol=n.Z)
dvd.beta<- matrix(rep(0,n.X*n.X), nrow=n.X, ncol=n.X)
dvy.gamma<- matrix(rep(0,n.Z), nrow=n.Z, ncol=1)
dvy.beta<- matrix(rep(0,n.X), nrow=n.X, ncol=1)

#i=1919

for(i in IDS)
{ index<-X$SID==i
  NumTeethi <-sum(index)
  Yi <- as.matrix(subset(Y,SID == i, c(-1)))
  Ri.gamma <- matrix(rep(alpha10, NumTeethi^2), nrow=NumTeethi)
  diag(Ri.gamma)<-1
  Vi.gamma <- diag(V.gamma[index])^(1/2) %*% Ri.gamma %*%
diag(V.gamma[index])^(1/2)
  dvd.gamma<-dvd.gamma+t(DZ[index,])%*% ginv( Vi.gamma) %*%DZ[index,]
  dvy.gamma<- dvy.gamma+t(DZ[index,])%*% ginv( Vi.gamma) %*%(ub[index,]-
p1[index])

  Ri.beta <- matrix(rep(alpha20, NumTeethi^2), nrow=NumTeethi)
  diag(Ri.beta)<-1
  Vi.beta<- diag(V.beta[index])^(1/2) %*% Ri.beta %*% diag(V.beta[index])^(1/2)
  dvd.beta<-dvd.beta+t(DX[index,])%*% ginv( Vi.beta) %*%diag(1-
ub[index,])%*%DX[index,]
  dvy.beta<- dvy.beta+t(DX[index,])%*% ginv( Vi.beta) %*%((1-
ub[index,])*(Y1[index]-lamda1[index]))
}
  gamma1 <- gamma0 + solve(dvd.gamma) %*% dvy.gamma
  beta1<- beta0 + solve(dvd.beta) %*% dvy.beta

####      Step 4: Renew tao      ####

```

```

lamda.new <- exp(X1%*%beta1)
eta.new <- exp(Z1%*%gamma1)
p.new<-eta.new/(1+eta.new)
# num.phi<-sum((1-ub)^2* lamda.new*(1+tao0*lamda.new)*(Y1-lamda.new))
# den.phi<-sum((1-ub)^2* lamda.new^2*(1+tao0*lamda.new)^2)
# phi.new<-num.phi/den.phi
num.tao<-sum(lamda.new^2*(1-ub)^2*((Y1-lamda.new)^2-lamda.new))
den.tao<-sum((1-ub)^2*lamda.new^4)
tao.new<-max(0.1,num.tao/den.tao)

#### Step 5: Renew alpha ####

u.alpha1<-(ub-p.new)/sqrt(p.new*(1-p.new))
u.alpha2<-(1-ub)*(Y1-lamda.new)/sqrt(lamda.new*(1+tao.new*lamda.new))
num1 <- num2<-0
den1 <- den2<-0
n1.total<-n1.star<-n2.total<-n2.star<-0
#i=1919
for(i in IDS)
{ index<-X$SID==i
temp.u1<-u.alpha1[index]
temp.u2<-u.alpha2[index]
ub.i<-1-ub[index]
NumTeethi <-sum(index)
a1.num<-sum(temp.u1%*%t(temp.u1))-sum(temp.u1^2)
a1.den<-sum(temp.u1^2)
num1 <- num1+a1.num
den1 <- den1+a1.den
n1.star<-n1.star + NumTeethi*(NumTeethi-1)
n1.total<-n1.total+NumTeethi
a2.num<-sum(temp.u2%*%t(temp.u2))-sum(temp.u2^2)
a2.den<-sum(temp.u2^2)
num2 <- num2+a2.num
den2 <- den2+a2.den
n2.star<-n2.star + sum(ub.i%*%t(ub.i))-sum(ub.i^2)
n2.total<-n2.total+sum(ub.i^2)
}

alpha1<-(num1/n1.star)/(den1/n1.total)
alpha2<-(num2/n2.star)/(den2/n2.total)
iteration.list <- list(beta=beta1, gamma=gamma1, tao=tao.new, alpha1=alpha1,
alpha2=alpha2)
return(iteration.list)
}

#### The GEE.ZINB will use the result from the first iteration ####

```

```

GEE.ZINB<- function(X, Y, Z, beta0, gamma0, tao0=tao0,
                    alpha10=0.5, alpha20=0.5)
{ Results.1<-GEE.ZINB.Single.Iter(X=X, Y=Y, Z=Z, beta0, gamma0,
  tao0, alpha10, alpha20)
  beta1=Results.1$beta; gamma1=Results.1$gamma; tao1=Results.1$tao
  alpha1=Results.1$alpha1; alpha2=Results.1$alpha2
  iter<-1
  while (max(abs(beta1-beta0), abs(gamma1-gamma0))>0.0001 & iter<50 & tao1>0.05)
  { beta0<-beta1; gamma0<-gamma1; tao0<-tao1
    alpha10<-alpha1; alpha20<-alpha1
    result<-GEE.ZINB.Single.Iter(X=X, Y=Y, Z=Z, beta0, gamma0,
      tao0,alpha10, alpha20)
    beta1=result$beta; gamma1=result$gamma; tao1=result$tao
    alpha1=result$alpha1; alpha2=result$alpha2
    iter<-iter+1
    ## print(iter)
  }
  if(!(iter<50)) print("Number of iterations is larger than 50")
  if(!(tao1>0.05)) print("Small Tao")
  if(!(iter<50 & tao1>0.05)) return(list(Converge="Error"))

```

```

#####      Generate Sandwich Variance      #####

```

```

IDS <- unique(X$SID)
X1<-X[, c(-1)]; X1<-as.matrix(cbind(rep(1,length(X[,1])), X1))
Z1<-Z[, c(-1)]; Z1<-as.matrix(cbind(rep(1,length(Z[,1])), Z1))
n.X <- length((X1[,1]))
n.Z <- length((Z1[,1]))
DX<-X1; DZ<-Z1
Y1 <- as.vector(Y[, c(-1)])
lamda1 <- exp(X1%*%beta1)
eta1 <- exp(Z1%*%gamma1)
p1<-eta1/(1+eta1)

```

```

#####   Update Uij   #####

```

```

temp<-(1+tao1*lamda1)^(-(1/tao1))
ub<-ifelse(Y1==0,1,0)/(1+(1-p1)*temp/p1)
# sum(ub==0); sum(Y1!=0)
# ub.gamma<-ifelse(Y1==0,1,0)/(1+(1-p1)*temp/p1)^2*temp/p1^2
# ub.beta<-ifelse(Y1==0,1,0)/(1+(1-p1)*temp/p1)^2*(1-p1)/p1
# ub.beta<-ub.beta*(1+tao1*lamda1)^(-(1/tao1)-1)

```

```

#####   Update gamma and beta   #####

```

```

DevBeta <- lamda1

```

```

DevGamma <- p1*(1-p1)
DX<-RowbyRow(X1, DevBeta)
DZ<-RowbyRow(Z1, DevGamma)
# ub.dev.gamma<-RowbyRow(DZ, ub.gamma)
# ub.dev.beta<-RowbyRow(DX, ub.beta)
V.gamma<-p1*(1-p1)
V.beta<-lamda1*(1+tao1*lamda1)

B.gamma.gamma<-M.gamma.gamma<- matrix(rep(0,n.Z*n.Z), nrow=n.Z, ncol=n.Z)
B.gamma.beta<-M.gamma.beta<- matrix(rep(0,n.Z*n.X), nrow=n.Z, ncol=n.X)
B.beta.gamma<-M.beta.gamma<- matrix(rep(0,n.X*n.Z), nrow=n.X, ncol=n.Z)
B.beta.beta<-M.beta.beta<- matrix(rep(0,n.X*n.X), nrow=n.X, ncol=n.X)
B1<-M<-matrix(rep(0,(n.X+n.Z)*(n.X+n.Z)), nrow=n.X+n.Z, ncol=n.X+n.Z)
dvy.gamma<- matrix(rep(0,n.Z), nrow=n.Z, ncol=1)
dvy.beta<- matrix(rep(0,n.X), nrow=n.X, ncol=1)

for(i in IDS)
{ index<-X$SID==i
  NumTeethi <-sum(index)
  Yi <- as.matrix(subset(Y,SID == i, c(-1)))
  Ri.gamma <- matrix(rep(alpha1, NumTeethi^2), nrow=NumTeethi)
  diag(Ri.gamma)<-1
  Vi.gamma <- diag(V.gamma[index])^(1/2) %*% Ri.gamma %*%
diag(V.gamma[index])^(1/2)
  B.gamma.gamma<-B.gamma.gamma+t(DZ[index,])%*%
ginv( Vi.gamma) %*%DZ[index,]
  # B.gamma.beta<-B.index,])%*% ginv( Vi.gamma) %*%(ub[index,]-p1[index])
  Ri.beta <- matrix(rep(alpha2, NumTeethi^2), nrow=NumTeethi)
  diag(Ri.beta)<-1
  Vi.beta<- diag(V.beta[index])^(1/2) %*% Ri.beta %*% diag(V.beta[index])^(1/2)
  # B.beta.gamma<-B.beta.gamma+t(DX[index,])%*%
ginv(Vi.beta)%*%diag(Y1[index]-lamda1[index])%*%ub.dev.gamma[index,]
  B.beta.beta<-B.beta.beta+t(DX[index,])%*%sqrt(diag(1-ub[index]))%*%
ginv(Vi.beta)%*%sqrt(diag(1-ub[index]))%*%DX[index,]
  dvy.gamma<-t(DZ[index,])%*% ginv( Vi.gamma) %*%(ub[index,]-p1[index])
  dvy.beta<- t(DX[index,])%*%ginv(Vi.beta) %*%diag(1-
ub[index])%*%(Y1[index]-lamda1[index])
  M[1:n.X, 1:n.X]<-M[1:n.X, 1:n.X]+dvy.beta%*%t(dvy.beta)
  M[1:n.X, (n.X+1):(n.X+n.Z)]<-M[1:n.X,
(n.X+1):(n.X+n.Z)]+dvy.beta%*%t(dvy.gamma)
  M[(1+n.X):(n.X+n.Z), 1:n.X]<-M[(1+n.X):(n.X+n.Z),
1:n.X]+dvy.gamma%*%t(dvy.beta)
  M[(1+n.X):(n.X+n.Z), (n.X+1):(n.X+n.Z)]<-M[(1+n.X):(n.X+n.Z),
(n.X+1):(n.X+n.Z)]+dvy.gamma%*%t(dvy.gamma)
}
B1[1:n.X, 1:n.X]<-B.beta.beta; B1[1:n.X, (n.X+1):(n.X+n.Z)]<-B.beta.gamma

```

```

      B1[(1+n.X):(n.X+n.Z), 1:n.X]<-B.gamma.beta; B1[(1+n.X):(n.X+n.Z),
(n.X+1):(n.X+n.Z)]<-B.gamma.gamma
      Var.model<-solve(B1)
      Var.Sand<-Var.model%%M%%Var.model
      se.beta<-sqrt(diag(Var.Sand))[1:n.X]
      se.gamma<-sqrt(diag(Var.Sand))[(n.X+1):(n.X+n.Z)]
      beta.fit<-cbind(beta=beta1, se=se.beta,
        z.test=beta1/se.beta, p.value=2*pnorm(-abs(beta1/se.beta)))
      gamma.fit<-cbind(gamma=gamma1, se=se.gamma,
        z.test=gamma1/se.gamma, p.value=2*pnorm(-abs(gamma1/se.gamma)))
      colnames(beta.fit)<-colnames(gamma.fit)<-c("Estimate", "S.E.", "Z-stat", "p-value")
      Result.Summary=list(Converge="YES", Beta=beta.fit, Gamma=gamma.fit,
        tao=tao1, alpha1=alpha1, alpha2=alpha2)
      return(Result.Summary)
    }

```

```

#####
##                                END                                ##
#####

```

```

#####
##                                Initial the GEE.ZINB                                ##
#####

```

```

####   Delete the Missing Value   ####

```

```

ok <- complete.cases(Dental)
sum(!ok); sum(ok)
NoMissing.Dental <- Dental[ok, ]
NoMissing.Dental$GenderM<-ifelse(NoMissing.Dental$Gender=="M", 1,0)

```

```

X <- subset(NoMissing.Dental, select=c(SID, GenderM, DentalExamAge,
      AUCmgF0_5yrs, AUCSodaOz0_5yrs,
      ToothBrushingFreqPerDayAvg, DentalVisitPast6moAvg,
      FluorideTreatment6moAvg, HomeFluorideppmAvg))
Z <- subset(NoMissing.Dental, select=c(SID, GenderM, DentalExamAge,
      AUCmgF0_5yrs, AUCSodaOz0_5yrs,
      ToothBrushingFreqPerDayAvg, DentalVisitPast6moAvg,
      FluorideTreatment6moAvg, HomeFluorideppmAvg))

```

```

Y <- subset(NoMissing.Dental, select=c(SID, CariesCount))

```

```

####   Initial GEE.ZINB by ZINB   ####

```

```

mZINB <- zeroinfl(formula=CariesCount~ Gender + DentalExamAge + AUCmgF0_5yrs
+ AUCSodaOz0_5yrs + ToothBrushingFreqPerDayAvg

```

```

+ DentalVisitPast6moAvg + FluorideTreatment6moAvg +
HomeFluorideppmAvg,
  dist = "negbin", data = NoMissing.Dental)
summary(mZINB)
names(summary(mZINB))

beta00<-as.vector(summary(mZINB)$coefficients$count[,1])
beta0<-beta00[c(-length(beta00))]
gamma0<-as.vector(summary(mZINB)$coefficients$zero[,1])
tao0<-exp(beta00[c(length(beta00))])
alpha10<-alpha20<-0.2

Result1<-GEE.ZINB(X, Y, Z, beta0, gamma0, tao0, alpha10=0.5, alpha20=0.5)
Result1

#####
##                               Reduced Model
#####

my.data<-NoMissing.Dental
X <- data.frame(SID=as.numeric(my.data$SID),
  DentalExamAge=as.numeric(my.data$DentalExamAge),
  AUCmgF0_5yrs=as.numeric(my.data$AUCmgF0_5yrs),
  DentalVisitPast6moAvg=as.numeric(my.data$DentalVisitPast6moAvg),
  FluorideTreatment6moAvg=as.numeric(my.data$FluorideTreatment6moAvg))
Z <- data.frame(SID=as.numeric(my.data$SID),

ToothBrushingFreqPerDayAvg=as.numeric(my.data$ToothBrushingFreqPerDayAvg))

Y <- data.frame(SID=as.numeric(my.data$SID),
  CariesCount=as.numeric(my.data$CariesCount))

mZINB <- zeroinfl(formula=CariesCount~ DentalExamAge + AUCmgF0_5yrs
+ DentalVisitPast6moAvg +
FluorideTreatment6moAvg|ToothBrushingFreqPerDayAvg,
  dist = "negbin", data = NoMissing.Dental)
summary(mZINB)

beta00<-as.vector(summary(mZINB)$coefficients$count[,1])
beta0<-beta00[c(-length(beta00))]
gamma0<-as.vector(summary(mZINB)$coefficients$zero[,1])
tao0<-exp(beta00[c(length(beta00))])
alpha10<-alpha20<-0.2
Result3<-GEE.ZINB(X, Y, Z, beta0,
  gamma0, tao0, alpha10=0.5, alpha20=0.5)
Result3

```

```
mGEEglm2<-geeglm(CariesCount~ DentalExamAge + AUCmgF0_5yrs  
+ DentalVisitPast6moAvg + FluorideTreatment6moAvg,  
corstr = "exchangeable", family = poisson, id = SID,data =NoMissing.Dental)  
summary(mGEEglm2)
```

```
#####
```

```
END
```

```
####
```

APPENDIX C

R code for Bootstrap Variance

```
#install.packages("geepack")
library(geepack)
ok <- complete.cases(Dental)
sum(!ok); sum(ok)
NoMissing.Dental <- Dental[ok, ]
NoMissing.Dental$GenderM<-ifelse(NoMissing.Dental$Gender=="M", 1,0)

X <- subset(NoMissing.Dental, select=c(SID, GenderM, DentalExamAge,
    AUCmgF0_5yrs, AUCSodaOz0_5yrs,
    ToothBrushingFreqPerDayAvg, DentalVisitPast6moAvg,
    FluorideTreatment6moAvg, HomeFluorideppmAvg))
Z <- subset(NoMissing.Dental, select=c(SID, GenderM, DentalExamAge,
    AUCmgF0_5yrs, AUCSodaOz0_5yrs,
    ToothBrushingFreqPerDayAvg, DentalVisitPast6moAvg,
    FluorideTreatment6moAvg, HomeFluorideppmAvg))

Y <- subset(NoMissing.Dental, select=c(SID, CariesCount))
#install.packages("geepack")
##library(geepack)
mZINB <- zeroinfl(formula=CariesCount~ Gender + DentalExamAge + AUCmgF0_5yrs
+ AUCSodaOz0_5yrs + ToothBrushingFreqPerDayAvg
+ DentalVisitPast6moAvg + FluorideTreatment6moAvg +
HomeFluorideppmAvg,
dist = "negbin", data = NoMissing.Dental)
summary(mZINB)
names(summary(mZINB))

beta00<-as.vector(summary(mZINB)$coefficients$count[,1])
beta0<-beta00[c(-length(beta00))]
gamma0<-as.vector(summary(mZINB)$coefficients$zero[,1])
tao0<-exp(beta00[c(length(beta00))])
alpha10<-alpha20<-0.2
beta.ZINB.est<-beta0
gamma.ZINB.est<-gamma0

Result1<-GEE.ZINB(X, Y, Z, beta0, gamma0, tao0, alpha10=0.5, alpha20=0.
Result1
beta.GEE.ZINB.est<-Result1$Beta[,1]
gamma.GEE.ZINB.est<-Result1$Gamma[,1]
```



```

mGEEglm <- geeglm(CariesCount~ Gender + DentalExamAge + AUCmgF0_5yrs +
AUCSodaOz0_5yrs + ToothBrushingFreqPerDayAvg
+ DentalVisitPast6moAvg + FluorideTreatment6moAvg +
HomeFluorideppmAvg,
corstr = "exchangeable", family = poisson, id = SID,data =NoMissing.Dental)
beta.GEE.est<-summary(mGEEglm)$coeff[,1]

Beta.GEE.ZINB<-Beta.GEE<-Beta.ZINB<-c()
Gamma.GEE.ZINB<-Gamma.ZINB<-c()

#### Bootstrap ####

set.seed(999)

#ID<-3
for (iter in (1:1000))
{ print(iter)
IDS <- unique(X$SID)
new.IDS<-sample(IDS, size=length(IDS), replace = TRUE)
new.data<-NoMissing.Dental[NoMissing.Dental$SID==new.IDS[1],]
new.data$SID<-1
for (my.id in 2:length(new.IDS))
{ index<-NoMissing.Dental$SID==new.IDS[my.id]
temp1<-NoMissing.Dental[index,]
temp1$SID<-my.id
new.data<-rbind(new.data, temp1)
}
XS <- subset(new.data, select=c(SID, GenderM, DentalExamAge,
AUCmgF0_5yrs, AUCSodaOz0_5yrs,
ToothBrushingFreqPerDayAvg, DentalVisitPast6moAvg,
FluorideTreatment6moAvg, HomeFluorideppmAvg))
ZS <- subset(new.data, select=c(SID, GenderM, DentalExamAge,
AUCmgF0_5yrs, AUCSodaOz0_5yrs,
ToothBrushingFreqPerDayAvg, DentalVisitPast6moAvg,
FluorideTreatment6moAvg, HomeFluorideppmAvg))
YS <- subset(new.data, select=c(SID, CariesCount))
mZINB <- zeroinfl(formula=CariesCount~ Gender + DentalExamAge +
AUCmgF0_5yrs + AUCSodaOz0_5yrs + ToothBrushingFreqPerDayAvg
+ DentalVisitPast6moAvg + FluorideTreatment6moAvg +
HomeFluorideppmAvg,
dist = "negbin", data = new.data)
beta00<-as.vector(summary(mZINB)$coefficients$count[,1])
beta0<-beta00[c(-length(beta00))]
gamma0<-as.vector(summary(mZINB)$coefficients$zero[,1])
Beta.ZINB<-rbind(Beta.ZINB,beta0)
Gamma.ZINB<-rbind(Gamma.ZINB, gamma0)

```

```

ResultS<-GEE.ZINB(X=XS, Y=YS, Z=ZS, beta0, gamma0, tao0, alpha10=0.5,
alpha20=0.5)
Beta.GEE.ZINB<-rbind(Beta.GEE.ZINB, ResultS$Beta[,1])
Gamma.GEE.ZINB<-rbind(Gamma.GEE.ZINB,ResultS$Gamma[,1])

mGEEglm <- geeglm(CariesCount~ Gender + DentalExamAge + AUCmgF0_5yrs +
AUCSodaOz0_5yrs + ToothBrushingFreqPerDayAvg
+ DentalVisitPast6moAvg + FluorideTreatment6moAvg +
HomeFluorideppmAvg,
corstr = "exchangeable", family = poisson, id = SID,data =new.data)
Beta.GEE<-rbind(Beta.GEE, summary(mGEEglm)$coeff[,1])
}

apply(Beta.GEE.ZINB,2, mean)
apply(Beta.ZINB,2, mean)
apply(Beta.GEE,2, mean)

apply(Beta.GEE.ZINB,2, sd)
apply(Beta.ZINB,2, sd)
apply(Beta.GEE,2, sd)

beta.GEE.ZINB.est
beta.GEE.ZINB.est/apply(Beta.GEE.ZINB,2, sd)
2*(1-pnorm(abs(beta.GEE.ZINB.est)/apply(Beta.GEE.ZINB,2, sd)))
beta.ZINB.est
beta.ZINB.est/apply(Beta.ZINB,2, sd)
2*(1-pnorm(abs(beta.ZINB.est)/apply(Beta.ZINB,2, sd)))
beta.GEE.est
beta.GEE.est/apply(Beta.GEE,2, sd)
2*(1-pnorm(abs(beta.GEE.est)/apply(Beta.GEE,2, sd)))

apply(Gamma.GEE.ZINB,2, mean)
apply(Gamma.ZINB,2, mean)

apply(Gamma.GEE.ZINB,2, sd)
apply(Gamma.ZINB,2, sd)

gamma.GEE.ZINB.est
gamma.GEE.ZINB.est/apply(Gamma.GEE.ZINB,2, sd)

2*(1-pnorm(abs(gamma.GEE.ZINB.est)/apply(Gamma.GEE.ZINB,2, sd)))
gamma.ZINB.est
gamma.ZINB.est/apply(Gamma.ZINB,2, sd)
2*(1-pnorm(abs(gamma.ZINB.est)/apply(Gamma.ZINB,2, sd)))

```

END

APPENDIX D

R code for Simulation

```
#####  
###                               Simulations  
#####  
  
c(length(X$SID), length(unique(X$SID)))  
  
####      8189 observations in 414 patients      ####  
  
install.packages("corcounts")  
library(corcounts)  
install.packages("bindata")  
library(bindata)  
install.packages("geepack")  
library(geepack)  
  
XZ<- subset(NoMissing.Dental, select=c(SID, DentalExamAge,  
    AUCmgF0_5yrs, DentalVisitPast6moAvg,  
    FluorideTreatment6moAvg, ToothBrushingFreqPerDayAvg))  
XZ1<-c()  
  
####      ID=3  
####      length(unique(XZ$SID))  
  
for (ID in unique(XZ$SID)[1:400])  
{ temp<-(XZ[XZ$SID==ID,])[1,]  
  for(j in 1:20)  
  { XZ1<-rbind(XZ1,temp)}  
}  
attributes(XZ1)  
XZ1.num<-data.frame(SID=as.numeric(XZ1$SID),  
    Intcept=rep(1, length(XZ1[,1])),  
    DentalExamAge=as.numeric(XZ1$DentalExamAge),  
    AUCmgF0_5yrs=as.numeric(XZ1$AUCmgF0_5yrs),  
    DentalVisitPast6moAvg=as.numeric(XZ1$DentalVisitPast6moAvg),  
    FluorideTreatment6moAvg=as.numeric(XZ1$FluorideTreatment6moAvg),  
  
    ToothBrushingFreqPerDayAvg=as.numeric(XZ1$ToothBrushingFreqPerDayAvg))  
X.sim<-as.matrix(XZ1.num[,2:6])  
Z.sim<-as.matrix(XZ1.num[,c(2,7)])  
##attributes(X.sim)
```

```

##beta.truth<-c(-3.51, 0.806, -0.936, 0.817, 1.222)
beta.truth<-c(-3.923, 0.935, -1.163, 0.565, 1.278)
lambda<-exp(X.sim%*%beta.truth)
##gamma.truth<-c(2.897, 0.513)
gamma.truth<-c(1.783, 0.551)
eta<-Z.sim%*%gamma.truth
##tao<-0.462
tao<-0.344
p.zero<-exp(eta)/(1+exp(eta))
unique(p.zero)
gencor.nb<-function(n, mu, tao,alpha)
{ X<-rnorm(n)
  R<-matrix(alpha, nrow=n, ncol=n)
  diag(R)<-1
  #t(chol(R))%*%chol(R)-R=0
  X.cor<-X%*%chol(R)
  x.nb<-qnbinom(p=pnorm(X.cor),size=ceiling(1/tao), prob=(1/tao)/(1/tao+mu))
  return(x.nb)
}
gencor.bi<-function(n, mu, alpha)
{ X<-rnorm(n)
  R<-matrix(alpha, nrow=n, ncol=n)
  diag(R)<-1
  #t(chol(R))%*%chol(R)-R=0
  X.cor<-X%*%chol(R)
  x.bi<-qbinom(p=pnorm(X.cor),size=1, prob=mu)
  return(x.bi)
}
## ID<-3
X<-XZ1.num[,c(-2, -7)]
Z<-XZ1.num[,c(1, 7)]

alpha0<-0.9
#alpha0<-0.5
#alpha0<-0.0
set.seed(999)
Beta.GEE.ZINB<-Beta.GEE<-Beta.ZINB<-c()
Beta.GEE.ZINB.var<-Beta.GEE.var<-Beta.ZINB.var<-c()
Beta.GEE.ZINB.mse<-Beta.GEE.mse<-Beta.ZINB.mse<-c()
Gamma.GEE.ZINB<-Gamma.ZINB<-c()
Gamma.GEE.ZINB.var<-Gamma.ZINB.var<-c()
Gamma.GEE.ZINB.mse<-Gamma.ZINB.mse<-c()

#ID<-3
for (iter in (1:1000))
{ resp.y<-c()

```

```

for (ID in unique(XZ1$SID)[1:400])
{ index<-XZ1$SID==ID
  n.obs<-sum(index)
  lambda.ID<-unique(lambda[index])
  p.zero.ID<-unique(p.zero[index])
  y.nb<-gencor.nb(n=n.obs, mu=lambda.ID, tao=tao, alpha=alpha0)
  y.zero<-gencor.bi(n=n.obs, mu=p.zero.ID, alpha=alpha0)
  y.obs<-ifelse(y.zero==1,0,y.nb)
  resp.y<-c(resp.y,y.obs)
}
Y<-data.frame(SID=XZ1.num$SID,CariesCount=resp.y)
Result2<-GEE.ZINB(X, Y, Z, beta0=beta.truth, gamma0=gamma.truth, tao0,
alpha10=0.5, alpha20=0.5)
if(Result2$Converge=="YES")
{ Beta.GEE.ZINB<-rbind(Beta.GEE.ZINB, Result2$Beta[,1])
  Beta.GEE.ZINB.var<-rbind(Beta.GEE.ZINB.var, Result2$Beta[,2]^2)
  Beta.GEE.ZINB.mse<-rbind(Beta.GEE.ZINB.mse, (Result2$Beta[,1]-beta.truth)^2)
  Gamma.GEE.ZINB<-rbind(Gamma.GEE.ZINB, Result2$Gamma[,1])
  Gamma.GEE.ZINB.var<-rbind(Gamma.GEE.ZINB.var, Result2$Gamma[,2]^2)
  Gamma.GEE.ZINB.mse<-rbind(Gamma.GEE.ZINB.mse, (Result2$Gamma[,1]-
gamma.truth)^2)
}
data.temp<-XZ1.num
data.temp$CariesCount<-resp.y
####data.temp[1:5,]
mZINB.temp <- zeroinfl(formula=CariesCount~DentalExamAge + AUCmgF0_5yrs
+ DentalVisitPast6moAvg +
FluorideTreatment6moAvg|ToothBrushingFreqPerDayAvg,
dist = "negbin", data = data.temp)
beta00<-as.vector(summary(mZINB.temp)$coefficients$count[,1])
beta0<-beta00[c(-length(beta00))]
gamma0<-as.vector(summary(mZINB.temp)$coefficients$zero[,1])
Beta.ZINB<-rbind(Beta.ZINB, beta0)
Beta.ZINB.var<-rbind(Beta.ZINB.var, summary(mZINB.temp)$coefficients$count[c(-
length(beta00)),2]^2)
Beta.ZINB.mse<-rbind(Beta.ZINB.mse, (beta0-beta.truth)^2)

Gamma.ZINB<-rbind(Gamma.ZINB, gamma0)
Gamma.ZINB.var<-rbind(Gamma.ZINB.var,
summary(mZINB.temp)$coefficients$zero[,2]^2)
Gamma.ZINB.mse<-rbind(Gamma.ZINB.mse,(gamma0-gamma.truth)^2)

mGEE <- geeglm(CariesCount~ DentalExamAge + AUCmgF0_5yrs
+ DentalVisitPast6moAvg + FluorideTreatment6moAvg,
corstr = "exchangeable", family = poisson, id = SID,data = data.temp)

```

```

Beta.GEE<-rbind(Beta.GEE,summary(mGEE)$coeff[,1])
Beta.GEE.var<-rbind(Beta.GEE.var,summary(mGEE)$coeff[,2]^2)
Beta.GEE.mse<-rbind(Beta.GEE.mse,(summary(mGEE)$coeff[,1]-beta.truth)^2)
}
beta.truth
alpha0
apply(Beta.GEE.ZINB,2, mean)
apply(Beta.ZINB,2, mean)
apply(Beta.GEE,2, mean)

apply(Beta.GEE.ZINB.var,2, mean)
apply(Beta.ZINB.var,2, mean)
apply(Beta.GEE.var,2, mean)

apply(Beta.GEE.ZINB.mse,2, mean)
apply(Beta.ZINB.mse,2, mean)
apply(Beta.GEE.mse,2, mean)

gamma.truth
apply(Gamma.GEE.ZINB,2, mean)
apply(Gamma.ZINB,2, mean)

apply(Gamma.GEE.ZINB.var,2, mean)
apply(Gamma.ZINB.var[!(is.na(Gamma.ZINB.var[,1])|is.na(Gamma.ZINB.var[,2])),],2,
mean)

apply(Gamma.GEE.ZINB.mse,2, mean)
apply(Gamma.ZINB.mse,2, mean)

#####
apply(Beta.GEE.ZINB,2, median)
apply(Beta.ZINB,2, median)
apply(Beta.GEE,2, median)

apply(Beta.GEE.ZINB.var,2, median)
apply(Beta.ZINB.var,2, median)
apply(Beta.GEE.var,2, median)

apply(Beta.GEE.ZINB.mse,2, median)
apply(Beta.ZINB.mse,2, median)
apply(Beta.GEE.mse,2, median)

gamma.truth
apply(Gamma.GEE.ZINB,2, median)
apply(Gamma.ZINB,2, median)

```

```
apply(Gamma.GEE.ZINB.var,2,median)
apply(Gamma.ZINB.var[!(is.na(Gamma.ZINB.var[,1])|is.na(Gamma.ZINB.var[,2])),],2,
median)
```

```
apply(Gamma.GEE.ZINB.mse,2, median)
apply(Gamma.ZINB.mse,2, median)
```

```
#####
####                                END
#####
```


CURRICULUM VITAE

NAME: Sheng Xu

ADDRESS: Department of Bioinformatics and Biostatistics
School of Public Health, University of Louisville
Louisville, KY 40292

E-MAIL: johnnysxu@gmail.com

EDUCATION:

B.S. Honors in Computer Science
Huaqiao University, Xiamen, China
September 2003-June 2007

Master of Public Administration
University of Baltimore, MD, U.S.
July 2009-December 2010

HONORS AND AWARDS:

- SAS Certified Base Programmer 2011
- SAS Certified Advanced Programmer 2011
- SAS Certified Clinical Trials Programmer 2012

RESEARCH

EXPERIENCE:

- Research experience
 1. University of Louisville, Louisville, KY
Research Assistant – NIH CAESAR study