2016

Analysis of Zero Inflated Over dispersed Count Data Regression Models with Missing Values

Mohammad Rajibul Islam Mian
University of Windsor

Analysis of Zero Inflated Over dispersed Count Data Regression Models with Missing Values

by

Mohammad Rajibul Islam Mian

A Dissertation
Submitted to the Faculty of Graduate Studies
through the Department of Mathematics and Statistics
in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy at the
University of Windsor

Windsor, Ontario, Canada

Analysis of Zero Inflated Over dispersed Count Data Regression Models with Missing Values

by

Mohammad Rajibul Islam Mian

APPROVED BY:

Dr. Leilei Zeng, External Examiner
University of Waterloo

Dr. Y. Aneja
Odette School of Business

Dr. M. Hlynka
Department of Mathematics and Statistics

Dr. A. Hussein
Department of Mathematics and Statistics

Dr. S. R. Paul, Supervisor
Department of Mathematics and Statistics

August 12, 2016

Author's Declaration of Originality

I hereby certify that I am the sole author of this thesis and that no part of this thesis has been published or submitted for publication.

I certify that, to the best of my knowledge, my dissertation does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my dissertation, published or otherwise, are fully acknowledged in accordance with standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained written permission from the copyright owner(s) to include such material(s) in my dissertation and have included copies of such copyright clearances to the appendix.

I declare that this is a true copy of my dissertation, including any final revisions, as approved by my dissertation committee and the Graduate Studies Office, and that this dissertation has not been submitted for a higher degree to any other University or Institution.

Abstract

Discrete data in the form of counts arise in many health science disciplines such as biology and epidemiology. The Poisson distribution is the most commonly used distribution for analysing count data. The Poisson distribution has a property that mean and the variance of the distribution are equal to each other. However, in many count data cases this property of the Poisson distribution does not hold as extra dispersion (variation) is observed in the data, and thus Poisson distribution is not an ideal choice for analysing count data in many applications. The presence of extra dispersion in count data is common in many real life situations. To accommodate this extra dispersion situation in count data a well known model is the negative binomial distribution, which is very convenient and common in practice. Often times a particular count (for example zero) may arise more than the expected number in the data. Count data with many zeros may not be explained properly by a model such as a Poisson distribution and a negative binomial distribution, so a zero inflated Poisson distribution and a zero inflated negative binomial distribution can be the ideal choice. Count data in the presence of both extra dispersion as well as zero inflation can be analysed by a zero inflated negative binomial model. Regression analysis of count data may be further complicated by the existence of missing values either in the response variable and/or in the explanatory variables (covariates).

In this dissertation we develop an estimation procedure for the parameters of the count data regression model with extra dispersion and zero inflation in the presence of missing values (1) in the response variable, (2) in the explanatory variables and

(3) both in the response and explanatory variables. We specifically use the extended negative binomial model as a count data model and address all three missing data mechanisms. A weighted expectation maximization algorithm (Ibrahim (1990)) is developed for the Maximum likelihood (ML) estimation of the parameters involved. Some simulations are conducted to study the properties of the estimates. Robustness of the procedure is shown when count data follow other over-dispersed models, such as the log-normal mixture of the Poisson distribution. An illustrative example using the dental epidemiology data of Bohning et al. (1999) and a discussion leading to some conclusions are given.

Dedication

Dedicated to my family.

Acknowledgements

I am very grateful to my creator Almighty ALLAH for His countless blessings over me and my family in every part of our life..

I would like to express my sincere gratitude to my supervisor Dr. Sudhir R. Paul for his supervision, guidance, encouragement and financial support during the preparation of this dissertation.

I am very thankful Dr. M. Hlynka, Dr. A. Hussein and Dr. Y. Aneja for being part of my Doctoral committee, and more importantly for their critical review and constructive suggestions and comments, which helped me to improve this dissertation. I would like to express my sincere gratitude to the external examiner of my dissertation Dr. Leilei Zeng, Department of Statistics and Actuarial Science, University of Waterloo for her critical review, constructive suggestions and comments.

I am very thankful to the Department of Mathematics and Statistics, UWindsor for providing me with the financial support such as Graduate Assistantships, Ontario Graduate Scholarship (OGS) and more importantly the position of Sessional Instructor.

I am also very thankful to my family members for their support, patience, love and prayers during my study period.

Mohammad Rajibul Islam Mian

August 12, 2016, Windsor, Ontario, Canada.

Contents

List of Tables

Chapter 1

Introduction

Discrete data in the form of counts arise in many health science disciplines such as biology and epidemiology. For examples of discrete count data see Deng and Paul (2000, 2005), Bohning, D., Dietz, E., Schlattmann, P., Mendonca, L., and Kirchner, U. (1999), Anscombe (1949); Bliss and Fisher (1953); Bliss and Owen (1958); McCaughran and Arnold (1976); Margolin, Kaplan and Zeiger (1981); Ross and Preece (1985)), Manton, Woodbury and Stallard (1981). The Poisson distribution is the most commonly used distribution for analysing count data. The Poisson distribution has a property that mean and the variance of the distribution are equal to each other. However, in many count data cases this property of the Poisson distribution does not hold, as extra dispersion (variation) is observed in the data, and thus Poisson distribution is not an ideal choice for analysing count data in many applications. The presence of extra dispersion in count data is common in many real life situations. To accommodate this extra dispersion situation in count data a well known model is the negative binomial distribution, which is very convenient and common in practice. For the applications of the negative binomial distribution see for example

Engel (1984); Breslow (1984); Margolin et al. (1989); Lawless (1987); Manton et al. (1981). The negative binomial distribution has flexibility in its parameterization and has been used differently by different authors. For example, see Paul and Plackett (1978); Barnwal and Paul (1988); Paul and Banerjee (1998); Piegorsch (1990), Deng and Paul (2000, 2005). Often times a particular count (for example zero) may arise in the data more than the expected number. Count data with many zeros may not be explained properly by a model such as a Poisson distribution and a negative binomial distribution, so a zero inflated Poisson distribution and a zero inflated negative binomial distribution can be the ideal choice. For example see Deng and Paul (2000, 2005), Ridout, Demetrio and Hinde (1998), Williamson, Lin, Lyles and Hightower (2007). Count data in the presence of both extra dispersion as well as zero inflation can be analysed by a zero inflated negative binomial model. Extensive work has been done to fit zero-inflated and over-dispersed count data model to real life data. For example, see Ridout, Demetrio and Hinde (1998), Hinde and Demetrio (1998), Li, Lu, Park, Kim, Brinkley and Peterson (1999) , Hall (2000) , Lee, Wang and Yau (2001) , Wang, Lee, Yau and Carrivick (2003), Lord, Washington and Ivan (2005) , Jiang and Paul (2009), Cameron and Trivedi (2013). Also a lot of work has been done to test the presence of zero-inflation and/or over-dispersion. For example, see Mullahy (1997), Dean (1992), Greene (1994), Broek (1995), Deng and Paul (2000), Xie, He, and Goh (2001), Paul, Jiang, Rai and Balasooriya (2004), Williamson, Lin, Lyles and Hightower (2007).

An example of count data in the presence of both extra dispersion as well as zero inflation can be found in Bohning, Dietz, Schlattmann, Mendonca, and Kirchner (1999). Bohning et al. (1999) present a set of data on a prospective study of dental status represented by decayed, missing and filled teeth (DMFT) index of school chil-

dren from an urban area of Belo Horizonte (Brazil). DMFT index scores can range from 0 to 28 or 32 per individual. The tooth is considered as decayed, when a carious lesion or both carious lesion and a restoration are present. The tooth is considered as missing if the tooth has been extracted due to caries. If a temporary or permanent filling is present in the tooth, or the filling of the tooth is defective but not decayed, then the tooth is considered as a filled tooth. The total number of tooth of a person having these properties would be the DMFT index for the person. More details of DMFT index can be found in Cappelli and Mobley (2007). The DMFT index was observed for 797 children at the beginning and at the end of the study. For the purpose of illustration here we consider DMFT index observed at the beginning of the study which, when summarized in terms of index and its frequency, are (index, frequency): (0,172), (1,73), (2,96), (3,80), (4,95), (5,83), (6,85), (7,65), (8,48). The mean and the variance of these counts are 3.3237 and 6.6387, which show over-dispersion in the data. Further, the observed frequency of zeros is 172 as opposed to the expected frequency of $797 \times P(x = 0) = 797 \times (0.036010) = 28.71$ showing that the data are also zero-inflated under a Poisson model.

Regression analysis of count data may be further complicated by the existence of missing values either in the response variable and/or in the explanatory variables (covariates). Extensive work has been done on regression analysis of continuous response data with some missing covariates under normality assumption. See, for example, Rubin (1977), Little and Rubin (1987, 2002, 2014), Lipsitz and Ibrahim (1996), Ibrahim, Chen and Lipsitz (1999), Ibrahim, Chen, Lipsitz and Herring (2005), Sinha and Maiti (2007), Maiti and Pradhan (2009).

Some work on missing values has also been done on logistic regression analysis of binary data. See, for example, Ibrahim (1990), Lipsitz and Ibrahim (1996), Ibrahim

and Lipsitz (1996), Ibrahim, Chen and Lipsitz (1999), Ibrahim, Chen and Lipsitz (2001), Sinha and Maiti (2007), Maiti and Pradhan (2009).

Rubin (1977) and Little and Rubin (1987, 2002, 2014) discuss various missingness mechanisms. If the missingness does not depend on observed data, then the missing data are called missing completely at random (MCAR). If the missing data mechanism depends only on observed data, then the data are missing at random (MAR). The MAR is also known as ignorable missing That is, in this case, the missing data mechanism is ignored. If the missing data mechanism depends on both observed and unobserved data, that is, failure to observe a value depends on the value that would have been observed, then the data are said to be missing not at random (MNAR) in which case the missingness is nonignorable . For more detailed discussion on missing data mechanism see Ibrahim et al. (2005).

In this dissertation, we develop an estimation procedure for the parameters of the count data regression model with extra dispersion and zero inflation in presence of missing values (1) in the response variable, (2) in the explanatory variables and (3) both in the response and explanatory variables. We specifically use the extended negative binomial model as a count data model and address all three missing data mechanisms. A weighted expectation maximization algorithm (Ibrahim (1990)) is developed for the Maximum likelihood (ML) estimation of the parameters involved. Some simulations are conducted to study the properties of the estimates. Robustness of the procedure is shown when count data follow other over-dispersed models, such as the log-normal mixture of the Poisson distribution. An illustrative example using the dental epidemiology data of Bohning et al. (1999) and a discussion leading to some conclusions are given.

We begin Chapter 2 by reviewing some literature related to zero inflated over dispersed count data, missing values and maximum likelihood estimation by using weighted expectation maximization algorithm.

In Chapter 3, we develop an estimation procedure for the parameters of the count data regression model with extra dispersion and zero inflation in presence of missing values in the response variable, that is, we assume that the regression variables are completely observed. Results of a simulation study with an illustrative example and a discussion leading to some conclusions is given.

Chapter 4 shows the estimation procedure for the parameters of the count data regression model with extra dispersion and zero inflation in presence missing values in the covariates. Results of a simulation study with an illustrative example and a discussion leading to some conclusions is given.

In Chapter 5, we develop the estimation procedure for the parameters of the count data regression model with extra dispersion and zero inflation in presence of missing values simultaneously both in the response variable and in the explanatory variables. Results of a simulation study with an illustrative example and a discussion leading to some conclusions is given as well.

A summary of this dissertation with some concluding remarks as well as a plan for future study are given in Chapter 6.

There is repetition in the chapters because the chapters are intended for submission as separate papers.

Chapter 2

Preliminaries and Literature Review

2.1 Zero inflated Over dispersed Count data Model

2.1.1 Poisson Model

Let Y be the count which follows the Poisson distribution. The probability mass function for the Poisson distribution is

$$f(y;\mu) = \frac{e^{-\mu}\mu^{y}}{y!}, \tag{2.1}$$

where μ is the mean parameter. The mean and the variance of Poisson distribution are both μ.

2.1.2 Negative Binomial Model

Let Y be a negative binomial random variable with mean parameter μ and dispersion parameter c. Then, using the terminology of Paul and Plackett (1978), Y has probability mass function

$$f(y;\mu,c) = \quad = \quad \frac{\Gamma(y+c^{-1})}{y!\Gamma(c^{-1})}\left(\frac{c\mu}{1+c\mu}\right)^{y}\left(\frac{1}{1+c\mu}\right)^{c^{-1}}, \tag{2.2}$$

for $y = 0, 1, ...,\ \mu > 0$. Now, for a typical Y, $Var(Y) = \mu(1+\mu c)$ and $c > -1/\mu$. This is the extended negative binomial distribution of Prentice (1986) which takes account over-dispersion as well as under-dispersion. When $c = 0$, the variance of the $NB(\mu, c)$ distribution becomes that of the Poisson(μ) distribution. Moreover, it can be shown that the limiting distribution of the $NB(\mu, c)$ distribution, as $c \to 0$, is the Poisson(μ).

2.1.3 Zero Inflated Poisson Model

If one specific count (in particular zero) is more frequent in the data, then the zero inflated Poisson model would be an appropriate choice for the data. Following Lee, Wang and Yau (2001), the zero inflated Poisson model can be written as

$$f(y_i|x_i;\mu,\omega) = \begin{cases} \omega + (1-\omega)e^{-\mu} & \text{if y = 0,} \\ (1-\omega)\dfrac{e^{-\mu}\mu^{y}}{y!} & \text{if y > 0.} \end{cases} \tag{2.3}$$

The mean and the variance of zero inflated Poisson model are $E(Y) = (1-\omega)\mu$ and $Var(Y) = (1-\omega)\mu(1+\mu\omega)$ respectively.

2.1.4 Zero Inflated Negative Binomial Model

The zero-inflated negative binomial regression model (see Deng and Paul, 2005) can be written as

$$f(y_i|x_i;\mu,c,\omega) = \begin{cases} \omega + (1-\omega)(\frac{1}{1+c\mu})^{c^{-1}} & \text{if y} = 0, \\ (1-\omega)\dfrac{\Gamma(y+c^{-1})}{y!\Gamma(c^{-1})}\left(\dfrac{c\mu}{1+c\mu}\right)^{y}\left(\dfrac{1}{1+c\mu}\right)^{c^{-1}} & \text{if y} > 0 \end{cases} \tag{2.4}$$

with $E(Y) = (1-\omega)\mu$, and $Var(Y) = (1-\omega)\mu[1+(c+\omega)\mu]$, where ω is the zero-inflation parameter. We denote this distribution by $ZINB(\mu, c, \omega)$ distribution.

2.2 Techniques of handling missing values

Missing observations are very common in any kinds of data set especially in longitudinal studies. There are several different ways to handle the data having missing observations. Following Little and Rubin (1987, 2002), in this section we briefly describe some general procedures of handling missing values in the response variable or in the explanatory variables or both in the response and explanatory variables.

2.2.1 Missing data mechanism

It is very important to know why and how the observations are missing in a data set. Based on different features of missingness, a missing data mechanism can be divided into three parts, missing completely at random, missing at random and missing not at random.

Missing completely at random, MCAR

If the missingness does not depends on observed as well as unobserved observations then this type of missingness is known as missing completely at random. In this type of missingness, probability of missingness is same for all the observations. For example, if answering a question depends on the result of a head after tossing a fair coin, then missingness of that answer is completely random.

Missing at random, MAR

If the missingness of an observation only depends on observed observations then this type of missingness is known as missing at random. In MAR, probability of missingness depends only on available observations not on unobserved observations. For example, missing information about age or income may depend on other available information.

Missing not at random, MNAR

If the missingness depends on observed as well as unobserved observations then this type of missingness is known as missing not at random. In MNAR, probability of missingness depends on both observed and unobserved observations. Dropouts in the medical studies can be a good example of the MNAR. A person in a study may not like the previous results and may be worried about the future results of the study and dropped out.

2.2.2 Methods based on availability of information

In the missing data analysis it is important to decide about the information that whether it should be included or not in the model. Based on this fact few approaches are available.

Complete case analysis

Complete case analysis only considers those individuals or subjects for whom all required information is available. In this method subjects having one or more missing information would be discarded from the analysis. This method has some advantages like, any standard statistical software can be used for the analysis and interpretations of the results will be very straight forward. If the number of missing observations is quite high in the data set then deletion of the data may lose some important features of the data and the result in small sample size of the data may not allow any sustainable analysis. Complete case analysis sometimes is good or consistent under MCAR mechanism but it does not work well for the MAR, and MNAR mechanisms.

Available case analysis

In available case analysis all the available information is considered and no information is discarded. This approach is better than the complete case scenario due to considering more information in the analysis than the complete cases. This method is applicable only under MCAR mechanism. Under this method, different subject will have different amounts of information, which might affect the results of the analysis.

2.2.3 Imputation based procedures

Instead of deleting the subjects with missing observations, it is possible to keep all the subjects in the analysis by imputing the missing observations. Usually imputation of the missing observations is carried out by substituting the values based on the available or observed data. One of the basic advantages of this method is that it uses the complete set of data for the analysis and the missing observations are replaced using the available informations. Based on the way of using the observed values in imputation, imputation based techniques can be divided into few a categories.

Last value carried forward imputation

In this method, last observed value of a subject is carried over to the next missing observations. This method can be used in monotone as well as nonmonotone settings of missingness. In this technique, missing observations are substituted by the same subject's last observed information and it is assumed that the last condition would continue for the next unobserved measurements. This assumption is very strong and often does not work well. This method is sometimes used in clinical studies and often produces biased estimates of the parameter of interest. Though this method is not good, it helps to understand the pattern of the observations over time.

Imputation by related observation

Sometimes related observations plays a good role in case of imputing the missing values. It may happen in a study that the mother's age and educational status for a child is missing then the father's information can be used to fill the mothers missing

information. Sometimes missing information about income can be filled by the income of another person doing the same kind of job.

Imputation by unconditional mean

In this type of imputation procedure, the missing value of a subject is replaced by the average of the available information of the same variable but from different subjects. So, in this technique the available observations for the subjects will not be used for imputation of his or her missing values.

Imputation by conditional mean

This approach of imputation was discussed by Buck (1960) and Little and Rubin (1987). Following Molenberghs, G., Thijs, H. , Jansen, I., Beunckens, C., Kenward, M. G., Mallinckrodt, C., Carroll, R. J. (2004) conditional mean imputation can be explained by considering a single normal sample. The mean and the covariance matrix is calculated from the complete case of the data in the first step, and then in the second step, information from the first step is used to calculate the conditional mean from a regression of missing values of a subject conditional on the actual observations. Conditional mean from the second step was used to replace the missing value.

Hot deck imputation

Hot deck imputation procedure uses similar responding units from the sample to replace the missing observations. This technique is one of the commonly used techniques. For example, if the information about the total number of persons in a house-

hold is missing then that information would be replaced by the similar household of that area.

Cold deck imputation

In this imputation technique, missing observations are replaced with a constant value from the external sources like previous survey or study. Replacing missing values by using the cold deck technique may not give good statistical inference because the conditions of the current and the previous survey may not same.

Imputation by substitution

In this imputation technique, the missing observations or the nonresponses are substituted by the information from different sources or subjects which were not included initially in the survey. This method is usually used in the data collection stage of the survey. For example, if a previously selected subject was not found during the survey, then the information would be collected from another subject who was not selected initially in the survey.

Regression imputation

Regression imputation uses the predicted values from the regression model to replace the missing observations. In this method predicted values were obtained from the regression of the missing observations on the observed values of that unit. For example, if the height and the weight were measured from thirty students of a class and the weight of a student was missing, then the weights of the twenty nine students

would be regressed on the heights and the regression coefficients would be used for the prediction of the missing weight for that specific height.

Stochastic regression imputation

In the stochastic regression imputation technique, missing values will be replaced by the predicted values from the regression plus a residual. This residual would be included to incorporate the uncertainty of the predicted values.

Multiple imputation

In the multiple imputation technique (Rubin 1978, 1987) missing values are replaced by more than one value. This technique considers the uncertainty raised due to estimating the missing values. This is a kind of modeling technique which produces data that maintains the overall variability of the population. This technique also helps to calculate the variance of estimates. Data obtained from this technique also keep the relationship with the existing variables.

2.2.4 Weighting procedures

In sample surveys, not all the samples need to be simple random samples. In that case, the probability of being selected in the sample will not be the same for all the observations. Population weights can be defined as the inverse of the probability of being sampled. If π_i is the probability of being sampled, then the population weights, p_i would be $p_i = \frac{1}{\pi_i}$ and the sampling weights come from the division of the population weights by their mean. So, the sampling weights w_i would be $w_i = \frac{p_i}{\bar{p}}$. If y is the

variable, then the weighted mean would be $\bar{y} = \frac{\sum w_i y_i}{\sum w_i}$.

2.2.5 Model based approach

In the model based technique, a model was developed for partially missing data and inference of the model is done on the basis of likelihood under the model. For estimating the parameters of the likelihood, there exist some estimation techniques, maximum likelihood is one of them. Model based methods are flexible, and interpretations based on model based methods are quite realistic. Though model based techniques are not so easy to implement for all kinds of data sets, this technique gives better results than other techniques.

There are quite a few ways existing to apply the model based approach to missing data analysis. We will discuss the basic idea very briefly. Two models can be distinguished based on the factorization of the joint likelihood of response and the missing data indicator variable, one is the selection model and another one is the pattern mixture model. Following Little and Rubin (1987), these two models are based on two different frameworks of the joint distribution of the response, $\mathbf{Y}$ and the missingness indicator variable, $\mathbf{R}$ and they can be expressed as

selection model:

$$f(\mathbf{Y}, \mathbf{R}|\mathbf{X}, \mathbf{Z}, \omega_1, \omega_2) = f(\mathbf{Y}|\mathbf{X}, \mathbf{Z}, \omega_1) f(\mathbf{R}|\mathbf{Y}, \mathbf{X}, \omega_2),$$

pattern mixture model:

$$f(\mathbf{Y}, \mathbf{R}|\mathbf{X}, \mathbf{Z}, \psi_1, \psi_2) = f(\mathbf{Y}|\mathbf{R}, \mathbf{X}, \mathbf{Z}, \psi_1) f(\mathbf{R}|\mathbf{X}, \psi_2),$$

where $\mathbf{X}$, $\mathbf{Z}$ is the matrix of covariates of the fixed and the random effects respectively, ω's and ψ's represent the parameters of specific parts of the model and the missingness indicator $\mathbf{R}$ can be defined as

$$R_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \text{ is observed} \\ 0 & \text{otherwise,} \end{cases}$$

where i represents subjects ($i = 1, 2, ..., m$) and j indicates the occasions ($j = 1, 2, ..., n_i$) of the observations.

The First part of the selection model indicates the distribution of response given the covariates and the second part shows the missingness indicator of the response is function of responses as well as covariates. In the pattern mixture model, responses are grouped according to the missingness patterns of the data and then these groups are used for the modelling purpose. Just as with the selection model, in the pattern mixture model, first part shows the distribution of response given the covariates for the groups and the second part shows the missingness patterns of the response is the function of only covariates not the responses.

2.3 Comparisons of different approaches for handling missing data

Missing values are very common in any field of analysis. Handling missing values is not always straight forward. We have discussed very briefly about a few existing approaches for the analysis of data with missing observations.

Complete case analysis and the available case analysis only works nicely at the missing completely at random (MCAR) mechanism. Almost all statistical software can be used easily to apply these two methods in the data set. Though the interpretation from the complete case analysis is quite straight forward and easy to make, this method can lose many important features of the data set due to deleting subjects having missing values. Complete case analysis sometimes suffers from lack of reliability of interpretation because of small sample size. In the available case analysis, though it keeps subjects having missing values, it does not work better than the complete case analysis. Calculating variance components may give problems due to not having same amount of information in all subjects.

Imputation based analysis is often preferable compared to the complete case analysis and available case analysis due to complete data set. Like complete case analysis, imputation based analysis can be applied by any statistical software very easily. In most cases, this procedure requires the MCAR mechanism, which is not very common. The results from the imputation based methods are quite unreliable and it is very hard to distinguish between the situations where this method works nicely and where they do not. Method, like last value carried forward, may be very unrealistic in some settings. Very often imputation based methods need specific adjustments for acceptable point estimates and sometimes these methods are not capable to give correct precision estimators (Verbeke and Molenberghs, 2000).

Model based methods are flexible and there is no assumptions and adjustments like the other methods. This approach can work with a large data set and gives large sample estimates. These methods are sometimes hard to apply but always gives better results for interpretations. Among selection and pattern mixture modeling approach, pattern mixture model is convenient to apply and easy to interpret. Most existing

statistical software can easily work with the pattern mixture approach to the model. Though there are few packages available for the selection model, the distributional assumption of the conditional density very often creates computational hazards in application. Little (1993) argued that the pattern mixture model is more flexible in situations where the data are not missing completely at random and this model shows proximity to the way sample survey experts consider the nonresponse situation.

2.4 Estimation procedures for the Parameters

There are a few methods of estimation available for estimating the parameters of generalized linear model. Due to the advancement of computation power, maximum likelihood, multiple imputation, weighted estimating equations, Bayesian estimation techniques become popular for handling missing observation in model based estimation technique. These methods have been used mostly to estimate the parameters of Normal and Binomial models.

In the model based procedure, a parametric model can be easily specified for the variable with missing observations. In likelihood based estimation, the likelihood function often is factored based on the observed or unobserved observations. In this type of situation, maximum likelihood estimation technique is easily applicable to estimate the parameters. Moreover, maximum likelihood estimates can be used to estimate the variance components from the second derivative of the log likelihood. Newton Raphson (NR), Nelder Maid (NM) and similar algorithms are available to maximize the complete data loglikelihood function. EM algorithm by Dempster, Laird and Rubin (1977), Weighted EM algorithm by Ibrahim (1990) and other techniques with the help of maximizing algorithms (NR, NM) are available to find the

maximum likelihood estimate of the parameters. Multiple Imputation is another likelihood based approach, where a multiple complete data set is created by filling in the missing observations. These complete (imputed) data sets are then analysed or optimized to estimate the parameters for each complete data set. The estimates from the multiple data sets are then combined by averaging the estimates of the parameters. Detailed discussion on Multiple Imputation is available in Little and Rubin (2002). In many practical situations, likelihood based estimation may not be possible to find due to incorrect distributional assumptions. In this type of situation, weighted estimating equations can be used to estimate the parameters. More details about the weighted estimating equations in the presence of missing observations are available in Lipsitz, Ibrahim and Zhao (1999). Bayesian approach is another way of estimating the parameters of the model for the data having missing observations. In this approach, prior distributions are specified for all the parameters in the model. Distribution assumption for the variables having missing observations are also necessary under the Bayesian approach. Detailed discussion about this approach is available in Ibrahim, Lipsitz and Chen (2002). All these estimation techniques have been elaborately studied and compared in many different scenarios over the years. Application of any one of these techniques depends on the situation that needs to be addressed. There is no unique superiority of these techniques. More detailed discussion on this is available in Ibrahim, Chen, Lipsitz and Herring (2005). Available applications of these techniques are mostly limited in binary and normal variables. As our main focus is to estimate the parameters of a zero inflated over dispersed count data model in the presence of missing response, we have applied one (Maximum likelihood estimation using weighted EM algorithm) of these competitive methods to observe the compatibility as well as the performance of the estimation techniques. Application of the other methods of estimation in the count data setting will be addressed in timely

fashion at elsewhere.

Expectation Maximization (EM) algorithm by Dempster, Laird and Rubin (1977) has been used to find the maximum likelihood estimates of the regression parameters of the model for the data having incomplete or missing observations in the response or covariates. Dawid and Skene (1979) used EM algorithm for the maximum likelihood estimation of the observer error rates. Bock and Aitkin (1981) used the EM algorithm for marginal maximum likelihood estimation. Laird and Ware (1982) used the EM algorithm for the random effects model for the longitudinal data. Shumway and Stoffer (1982) used the EM algorithm for the time series modelling. Laird, Lange and Stram (1987) used EM algorithm for the maximum likelihood computations with the repeated measures. Lauritzen (1995) used EM algorithm for graphical association model. Ibrahim (1990) used the EM algorithm by the method of weights for the incomplete data in generalized linear models. Following Ibrahim (1990), series of articles have addressed the application of EM algorithm by the method weights. For more details, please see the following articles Lipsitz and Ibrahim (1996), Ibrahim and Lipsitz (1996), Ibrahim, Chen and Lipsitz (1999, 2001), Ibrahim, Chen, Lipsitz and Herring (2005), Sinha and Maiti (2007), Maiti and Pradhan (2009). EM algorithm by the method of weights is computationally more feasible and the implementation is straight forward. In the EM by the method of weights, log likelihood function for the parameters can be separated for the regression parameters, parameters of the covariate distribution and the parameters of the missingness mechanism. This feature of the log likelihood facilitates the separate maximization and helps to separate the nuisance parameters from the parameters of interest. These characteristics of the EM algorithm motivates us to use this algorithm to find the maximum likelihood estimates of the zero inflated over dispersed count data model with missing observations. More

details of this EM algorithm by the method of weights are explained in the following chapters.

Chapter 3

Estimation for Zero Inflated Over dispersed Count Data Model with Missing Response

3.1 Introduction

Discrete data in the form of counts often exhibit extra dispersion as well as zero inflation. For example, Bohning, Dietz, Schlattmann, Mendonca and Kirchner (1999) present a set of data on a prospective study of dental status measured by decayed, missing and filled teeth (DMFT) index of school children from an urban area of Belo Horizonte (Brazil), the Belo Horizonte caries prevention study. There were 797 children from six different schools who took part in the study. The children were all 7 years of age at the beginning of the study. Only the eight deciduous molars were considered so the smallest possible value of the DMFT index is 0 and the largest is

8. The prospective study was for a period of two years.

The caries prevention study was conducted to compare four methods of treatments of dental hygiene: oral health education, enrichment of the school diet with rice bran, mouthwash with 0.2% sodium fluoride solution, and oral hygiene. These four treatments along with no prevention measure (control) and all four methods together were randomized to the six schools.

The data then involved 3 categorical covariates: gender having two categories (0 - female, 1 - male), ethnic group having three categories (1 - dark, 2 - white, 3 - black) and school having six categories (1 - oral health education, 2 - all four methods together, 3 - control school (no prevention measure), 4 - enrichment of the school diet with rice bran, 5 - mouthwash with 0.2% NaF-solution, 6 - oral hygiene).

The DMFT index was obtained at the beginning of the study and also at the end of the study. For the purpose of illustration here we consider the DMFT index observed at the beginning of the study which, when summarized in terms of index and its frequency, are [(index, frequency): (0,172), (1,73), (2,96), (3,80), (4,95), (5,83), (6,85), (7,65), (8,48)]. The mean and the variance of these counts are 3.3237 and 6.6387, which show over-dispersion in the data. Further, the observed frequency of zeros is 172 as opposed to the expected frequency of 28.71 ($797 \times P(x = 0) = 797 \times (0.036010)$) under a Poisson model. These data thus show over-dispersion as well as zero inflation under a Poisson model.

A popular over-dispersed count data model is the two parameter negative binomial model. Different authors have used different parameterizations for the negative binomial distribution (see, for example, Paul and Plackett, 1978; Barnwal and Paul, 1988; Paul and Banerjee, 1998 and Piegorsch, 1990). Let Y be a negative binomial

random variable with mean parameter μ and dispersion parameter c. Then, using the terminology of Paul and Plackett (1978), Y has the probability mass function

$$f(y;\mu,c) = \quad = \quad \frac{\Gamma(y+c^{-1})}{y!\Gamma(c^{-1})}\left(\frac{c\mu}{1+c\mu}\right)^{y}\left(\frac{1}{1+c\mu}\right)^{c^{-1}}, \tag{3.1}$$

for $y = 0, 1, ..., \mu > 0$. Now, for a typical Y, $Var(Y) = \mu(1+\mu c)$ and $c > -1/\mu$. This is the extended negative binomial distribution of Prentice (1986) which takes account of over-dispersion as well as under-dispersion. Obviously, when $c = 0$, variance of the $NB(\mu, c)$ distribution becomes that of the Poisson(μ) distribution. Moreover, it can be shown that the limiting distribution of the $NB(\mu, c)$ distribution, as $c \to 0$, is the Poisson(μ).

Using the mass function in equation 3.1 the zero-inflated negative binomial regression model (see Deng and Paul, 2005) can be written as

$$f(y_i|x_i;\mu,c,\omega) = \begin{cases} \omega + (1-\omega)(\frac{1}{1+c\mu})^{c^{-1}} & \text{if y = 0,} \\ (1-\omega)\dfrac{\Gamma(y+c^{-1})}{y!\Gamma(c^{-1})}\left(\dfrac{c\mu}{1+c\mu}\right)^{y}\left(\dfrac{1}{1+c\mu}\right)^{c^{-1}} & \text{if y > 0} \end{cases} \tag{3.2}$$

with $E(Y) = (1-\omega)\mu$, and $Var(Y) = (1-\omega)\mu[1+(c+\omega)\mu]$, where ω is the zero-inflation parameter. We denote this distribution by $ZINB(\mu, c, \omega)$ distribution.

Extensive work has been done to fit zero-inflated and over-dispersed count data model to real life data. For example, see Cameron and Trivedi (1986), Dean (1992), Greene (1994), Broek (1995), Mullahy (1997), Ridout, Demetrio and Hinde (1998), Hinde and Demetrio (1998), Li et al. (1999), Hall (2000), Deng and Paul (2000, 2005), Lee et al. (2001), Xie, He and Goh (2001), Wang, Lee, Yau, and Carrivick

(2001), Paul, Jiang, Rai, and Balasooriya (2004), Williamson, Lin, Lord, Washington and Ivan (2005), Lyles and Hightower (2007), Jiang and Paul (2009), Synnott and Angers (2009). Also a lot of work has been done to test the presence of zero-inflation and/or over-dispersion. For example, see Cameron and Trivedi (1986), Dean (1992), Greene (1994), Broek (1995), Mullahy (1997), Ridout, Demetrio and Hinde (1998), Hinde and Demetrio (1998), Li et al. (1999), Hall (2000), Deng and Paul (2000, 2005), Lee et al. (2001), Xie, He and Goh (2001), Wang, Lee, Yau, and Carrivick (2001), Paul, Jiang, Rai, and Balasooriya (2004), Williamson, Lin, Lord, Washington and Ivan (2005), Lyles and Hightower (2007), Jiang and Paul (2009), Synnott and Angers (2009).

Regression analysis of count data may be further complicated by the existence of missing values either in the response variable and/or in the explanatory variables (covariates). Extensive work has been done on regression analysis of continuous response data with some missing responses under normality assumption. See, for example, Rubin (1977), Little and Rubin (1987), Anderson and Taylor (1976), Geweke (1986), Raftery, Madigan and Hoeting (1997), Chen, Hubbard and Rubin (2001), Kelly (2007), Zhang, C-H and Huang, J. (2008).

Some work on missing values has also been done on logistic regression analysis of discrete data. See, for example, Ibrahim (1990), Lipsitz and Ibrahim (1996), Ibrahim and Lipsitz (1996), Ibrahim, Chen and Lipsitz (1999, 2001), Ibrahim, Chen, Lipsitz and Herring (2005), Sinha and Maiti (2007), Maiti and Pradhan (2009).

Rubin (1977) and Little and Rubin (1987) discuss various missingness mechanisms. If the missingness does not depend on observed data, then the missing data are called missing completely at random (MCAR). If the missing data mechanism depends only

on observed data, then the data are missing at random (MAR). The MAR is also known as ignorable missing. That is, in this case, the missing data mechanism is ignored. If the missing data mechanism depends on both observed and unobserved data, that is, failure to observe a value depends on the value that would have been observed, then the data are said to be missing not at random (MNAR) in which case the missingness is nonignorable . For more detailed discussion on missing data mechanism, see Ibrahim et al. (2005, p333).

There are a few methods of estimation available for estimating the parameters of interest at the presence of missing values. Due to the advancement of computation power, maximum likelihood, multiple imputation, weighted estimating equations, Bayesian estimation techniques become popular for handling missing observation in model based estimation technique. These methods have been used mostly to estimate the parameters of Normal and Binomial models.

In the model based procedure, a parametric model can be easily specified for the variable with missing observations. In the likelihood based estimation, the likelihood function often factored based on the observed or unobserved observations. In this type of situation maximum likelihood estimation technique is easily applicable to estimate the parameters. Moreover, maximum likelihood estimates can be used to estimate the variance components from the second derivative of the log likelihood. Newton Raphson (NR), Nelder Maid (NM) and similar algorithms are available to maximize the complete data log likelihood function. EM algorithm by Dempster, Laird and Rubin (1977), Weighted EM algorithm by Ibrahim (1990) and other techniques with the help of maximizing algorithms (NR, NM) are available to find the maximum likelihood estimate of the parameters. Multiple Imputation is another likelihood based approach, where multiple complete data set being created by filling the the missing observa-

tions. These complete (imputed) data sets then analysed or optimized to estimate the parameters for each complete data set. The estimates from the the multiple data sets then combined by averaging the estimates of the parameters. Detailed discussion on Multiple Imputation are available in Little and Rubin (2002). In many practical situations, likelihood based estimation may not be possible to find due to incorrect distributional assumptions. In this type of situation weighted estimating equations can be used to estimate the parameters. More details about the weighted estimating equations in the presence of missing observations are available in Lipsitz, Ibrahim and Zhao (1999). Bayesian approach is another way of estimating the parameters of the model for the data having missing observations. In this approach, prior distributions are specified for all the parameters in the model. Distribution assumption for the variables having missing observations are also necessary under Bayesian approach. Detailed discussion about this approach is available in Ibrahim, Lipsitz and Chen (2002). All these estimation techniques have been elaborately studied and compared in many different scenarios over the years. Application of any one of these techniques depends on the situation that needs to be addressed. There is no unique superiority of these techniques. More detailed discussion on this is available in Ibrahim, Chen, Lipsitz and Herring (2005). Available applications of these techniques are mostly limited in binary and normal variables. As our main focus is to estimate the parameters of a zero inflated over dispersed count data model at the presence of missing response, we have applied one (Maximum likelihood estimation using weighted EM algorithm) of these competitive methods to observe the compatibility as well as the performance of the estimation techniques. Application of the other methods of estimation in the count data setting will be addressed in timely fashion at elsewhere.

Expectation Maximization (EM) algorithm by Dempster, Laird and Rubin (1977)

has been used to find the maximum likelihood estimates of the regression parameters of the model for the data having incomplete or missing observations in the response or covariates. Dawid and Skene (1979) used EM algorithm for the maximum likelihood estimation of the observer error rates. Bock and Aitkin (1981) used the EM algorithm for marginal maximum likelihood estimation. Laird and Ware (1982) used the EM algorithm for the random effects model for the longitudinal data. Shumway and Stoffer (1982) used the EM algorithm for the time series modelling. Laird, Lange and Stram (1987) used EM algorithm for the maximum likelihood computations with the repeated measures. Lauritzen (1995) used EM algorithm for graphical association model. Ibrahim (1990) used the EM algorithm by the method of weights for the incomplete data in generalized linear models. Following Ibrahim (1990), series of articles have addressed the application of EM algorithm by the method weights. For more details please see the following articles Lipsitz and Ibrahim (1996), Ibrahim and Lipsitz (1996), Ibrahim, Chen and Lipsitz (1999, 2001), Ibrahim, Chen, Lipsitz and Herring (2005), Sinha and Maiti (2007), Maiti and Pradhan (2009). EM algorithm by the method of weights is computationally more feasible and the implementation is straight forward. In the EM by the method of weights, log likelihood function for the parameters can be separated for the regression parameters, parameters of the covariate distribution and the parameters of the missingness mechanism. This feature of the log likelihood facilitates the separate maximization and helps to separate the nuisance parameters from the parameters of interest. These characteristics of the EM algorithm motivates us to use this algorithm to find the maximum likelihood estimates of the zero inflated over dispersed count data model with missing observations in the response. More details of this EM algorithm by the method of weights are explained in the following sections.

The purpose of this paper is to develop estimation procedure for the parameters of a count data regression model with extra dispersion and zero inflation in of presence missing values in the response variable. We specifically use the extended negative binomial model 3.1 as a count data model (which includes the Poisson regression model when the dispersion parameter tends to zero) and address inference problems under different missing data mechanism (MCAR, MAR and MNAR). The usual maximum likelihood estimation procedure becomes complicated under some missing data scenario as it involves multiple integration. So, following Ibrahim (1990) we develop a weighted expectation maximization (EM) algorithm. For various advantages of the EM algorithm as opposed to direct maximization in presence of missing values see the seminal paper by Dempster, Laird and Rubin (1977). Some simulations are conducted to study the properties of the estimators. A study of robustness of the procedure is also conducted for the situation when count data follow other over-dispersed models, such as the log-normal mixture of the Poisson distribution. The DMFT data discussed above are used to illustrate the procedure and a discussion is given.

The procedure for the estimation of the parameters are developed in Section 2. Results of a simulation study is reported in Section 3. An illustrative example using the dental epidemiology data of Bohning et al. (1999) is given in Section 4 and a discussion leading to some conclusions is given in Section 5.

3.2 Estimation in Zero-inflated and over-dispersed count data regression model with missing values in the Response Variable

Suppose data for the i^{th} of n subjects are (y_i, x_i), $i = 1, \ldots, n$, which are realizations from $ZINB(\mu, c, \omega)$, where y_i represents the response variable and x_i represents a $p \times 1$ vector of covariates with the regression parameter $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$, such that $\mu_i = \exp(\sum_{j=1}^{p} X_{ij}\beta_j)$. Here β_1 is the intercept parameter in which case $X_{i1} = 1$ for all i.

3.2.1 Estimation of the parameters with no missing data

For complete data, the likelihood function is

$$\begin{aligned} L(\beta, c, \omega | y_i) = \prod_{i=1}^{n} \Big[& (\omega + (1-\omega) f(0; \mu_i, c, \omega)) I_{\{y_i = 0\}} \\ & + (1-\omega) f(y_i; \mu_i, c, \omega) I_{\{y_i > 0\}} \Big]. \end{aligned} \tag{3.3}$$

Writing $\gamma = \omega/(1-\omega)$, so $\omega = \gamma/(1+\gamma)$ and $1 - \omega = 1/(1+\gamma)$ the log likelihood,

apart from a constant, can be written as

$$\begin{aligned} l(\beta, c, \gamma | y_i) = & \sum_{i=1}^{n} \Big[-\log(1+\gamma) + \log[\gamma + f(0; \mu_i, c, \omega)] I_{\{y_i=0\}} \\ & + \log f(y_i; \mu_i, c, \omega) I_{\{y_i>0\}} \Big] \\ = & \sum_{i=1}^{n} \Bigg[-\log(1+\gamma) + \log\big[\gamma + \exp[-c^{-1}\log(1+\mu_i c)]\big] I_{\{y_i=0\}} \\ & + \Big[(y_i \log \mu_i - (y_i + c^{-1}) \log(1+\mu_i c) + \sum_{l=1}^{y_i} [1 + (l-1)c]) \Big] I_{\{y_i>0\}} \Bigg]. \end{aligned} \tag{3.4}$$

The parameters β_j, c and γ can be estimated by directly maximizing the loglikelihood function 3.4 or by simultaneously solving the following estimating equations

$$\begin{aligned} \frac{\partial l}{\partial \beta_j} = & \sum_{i=1}^{n} \Bigg[\Big[\frac{-(1+\mu c)^{-1} \exp[(-c^{-1}\log(1+\mu c))]}{\gamma + \exp[(-c^{-1}\log(1+\mu c))]} I_{\{y_i=0\}} \\ & + \big[\frac{y_1}{\mu} - \frac{c(y_1 + c^{-1})}{1+\mu c} \big] I_{\{y_i>0\}} \Big] \frac{\partial \mu_i}{\partial \beta_j} \Bigg] = 0, \end{aligned} \tag{3.5}$$

$$\begin{aligned} \frac{\partial l}{\partial c} = \sum_{i=1}^{n} \Bigg[& \frac{[-\mu c^{-1}(1+\mu c)^{-1} + c^{-2}\log(1+\mu c)] \exp[(-c^{-1}\log(1+\mu c))]}{\gamma + \exp[(-c^{-1}\log(1+\mu c))]} I_{\{y_i=0\}} \\ & + \big[\mu(y_i + c^{-1})(1+\mu c)^{-1} - c^{-2}\log(1+\mu c) + \sum_{l=1}^{y_i} (l-1) \big] I_{\{y_i>0\}} \Bigg] = 0, \end{aligned} \tag{3.6}$$

and

$$\frac{\partial l}{\partial \gamma} = \sum_{i=1}^{n} \Big[-(1+\gamma)^{-1} + \big[\gamma + \exp[(-c^{-1}\log(1+\mu c))]\big]^{-1} I_{\{y_i=0\}} + 0 I_{\{y_i>0\}} \Big] = 0, \tag{3.7}$$

where $\dfrac{\partial \mu_i}{\partial \beta_j} = X_{ij} \exp(\sum_{j=1}^{p} X_{ij}\beta_j)$.

3.2.2 Estimation of the parameters with missing responses

Estimation under MCAR

In case of MCAR, missingness of the data do not depend on observed data and the subjects having the missing observations are deleted before the analysis. For estimation procedure the log likelihood function remains the same as equation 3.4 with reduced sample size having only complete observations.

Estimation under MAR

As some of the observations in response may be missing we write the response y_i as

$$y_i = \begin{cases} y_{o,i} & \text{if } y_i \text{ is observed,} \\ y_{m,i} & \text{if } y_i \text{ is missing.} \end{cases} \tag{3.8}$$

Using this in $f(y_i|x_i;\mu,c,\omega)$ given in equation 3.2, the log-likelihood is

$$\begin{aligned} l(\psi|Y_o,Y_m,X) = & \sum_{i=1}^{n} \log(f(y_i|x_i,\psi)) \\ = & \sum_{i=1}^{n} \Big[-\log(1+\gamma) + \log[\gamma + f(0;\mu_i,c,\omega)] I_{\{y_i=0\}} \\ & + \log f(y_i;\mu_i,c,\omega) I_{\{y_i>0\}} \Big], \end{aligned} \tag{3.9}$$

where Y_o is the vector of observed values, Y_m is the vector of missing values, $\psi = (\beta, c, \gamma)$ and $\mu_i = \exp(\sum_{j=1}^{p} X_{ij}\beta_j)$.

In MAR, conditional probability of missingness of the data depends on observed data. Parameters of the missingness mechanism are completely separate and distinct

from the parameters of the model 3.2. In likelihood based estimation considering MAR, missingness mechanism can be ignored from the likelihood and missing data that are missing at random are often known as ignorable missing or ignorable non-response, but the subjects having these missing observations cannot be deleted before the analysis (see Little and Rubin (1987), Ibrahim et al. (2005) for detailed discussion on this).

In this scenario, our goal is to maximize the following log likelihood with respect to the parameters ψ

$$l(\psi|Y_o, X) = \sum_{Y_m} l(\psi|Y_o, Y_m, X). \tag{3.10}$$

In the most general case where missing data are not MAR, the missing data process has to be modeled and included in the observed likelihood construction.

Direct maximization of $l(\psi; Y_o, X)$ is not, in general, straight forward. However, the EM algorithm (Dempster, Laird and Rubin (1977)) is a very useful tool for obtaining maximum likelihood estimates with missing observations.

The EM algorithm uses two iterative steps known as the expectation-step (E-step) and the maximization-step (M-step). Following Little and Rubin (1987), the E-step provides the conditional expectation of the log-likelihood $l(\psi|Y_o, Y_m, X)$ given the observed data (Y_o, X) and current estimate of the parameters ψ.

Suppose A of the n responses are observed and $B = n - A$ responses are missing and let s be an arbitrary number of iterations during maximization of the log-likelihood. Then the E-step of the EM algorithm for the i^{th} missing response for the $(s+1)^{th}$

iteration can be written as

$$\begin{aligned} Q_i(\psi|\psi^{(s)}) &= E[l(\psi|y_{o,i}, y_{m,i}, x_i)|y_{o,i}, x_i, \psi^{(s)}] \\ &= \sum_{y_{m,i}} l(\psi|y_{o,i}, y_{m,i}, x_i) P(y_{m,i}|x_i, \psi^{(s)}). \end{aligned} \tag{3.11}$$

For all the observations, the E-step of EM algorithm for the $(s+1)^{th}$ iteration is

$$Q(\psi|\psi^{(s)}) = \sum_{i=1}^{A} l(\psi|y_{o,i}) + \sum_{i=1}^{B} \sum_{y_{m,i}} l(\psi|y_{m,i}, x_i) P(y_{m,i}|x_i, \psi^{(s)}). \tag{3.12}$$

Note for the situation in which there is no missing response, the EM algorithm requires only maximization of the first term on the right hand side.

Here $P(y_{m,i}|x_i, \psi^{(s)})$ is the conditional distribution of the missing response given the observed data and the current (s^{th} iteration) estimate of ψ. However, in many situations, $P(y_{m,i}|x_i, \psi^{(s)})$ may not always be available. Following Ibrahim et al. (2001) and Sahu and Roberts (1999), we can write $P(y_{m,i}|x_i, \psi^{(s)}) \propto P(y_i|x_i, \psi^{(s)})$ (the complete data distribution given in 3.2). For the i^{th} of the B missing responses we take a sample $a_{i1}, a_{i2}, ..., a_{im_i}$ from $P(y_i|x_i, \psi^{(s)})$ using the Gibbs sampler (see Casella and George (1992) for details). Then, following Ibrahim et al. (2001) $Q(\psi|\psi^{(s)})$ can be written as

$$Q(\psi|\psi^{(s)}) = \sum_{i=1}^{A} l(\psi|y_{o,i}) + \sum_{i=1}^{B} \frac{1}{m_i} \sum_{k=1}^{m_i} l(\psi|x_i, a_{ik}). \tag{3.13}$$

In the M-step of the EM algorithm, the $Q(\psi|\psi^{(s)})$ is maximized. Here maximizing $Q(\psi|\psi^{(s)})$ is analogous to maximization of complete data log likelihood where each incomplete response is replaced by m_i weighted observations. More details of the

EM algorithm by the method of weights can be found in Ibrahim (1990), Lipsitz and Ibrahim (1996), Ibrahim and Lipsitz (1996), Ibrahim et al. (1999, 2001), Ibrahim et al.(2005), Sinha and Maiti (2007), Maiti and Pradhan (2009).

The variance covariance matrix of the estimates of the parameters is calculated by inverting the observed information matrix at convergence (Efron and Hinkley, 1978) which is

$$H_{\psi\psi'} = Q''(\psi|\psi^{(s)}) = \sum_{i=1}^{A} \frac{\partial^2}{\partial\psi\partial\psi'} l(\psi|y_{o,i}) + \sum_{i=1}^{B} \frac{1}{m_i} \sum_{k=1}^{m_i} \frac{\partial^2}{\partial\psi\partial\psi'} l(\psi|x_i, a_{ik}). \quad (3.14)$$

Expressions for the elements of H above are given in the Appendix.

Estimation under MNAR

Under MNAR, the probability of missing observations in the response variable depends on the the covariates and the values of the response that would have been observed. This missing data mechanism cannot be ignored and needs to be incorporated in the likelihood. The missing observations that follow this missing data mechanism are known as nonignorable missing. It is then necessary to specify a parametric model for this missingness. To put things in perspective, define a random variable $r_i (i = 1, 2, \ldots, n)$ as

$$r_i = \begin{cases} 0 & \text{if } y_i \text{ is observed,} \\ 1 & \text{if } y_i \text{ is missing.} \end{cases} \quad (3.15)$$

The random variable r_i follows

$$p(r_i|y_i, x_{ij}) = [p(r_i = 1)]^{r_i}[1 - p(r_i = 1)]^{(1-r_i)}. \tag{3.16}$$

See Ibrahim, Chen and Lipsitz (2001). Then, using a logit link function

$$\log[\frac{p(r_i = 1)}{1 - p(r_i = 1)}] = \nu_0 + \nu_1 * y_i + \nu_2 * x_{i1} + \nu_3 x_{i2} + \ldots + \nu_q x_{ip}, \tag{3.17}$$

where y_i is the responses and the responses that would have been observed, $x_{ij}(j = 1, 2, \ldots, p)$ are the covariates and $\nu = (\nu_0, \nu_1, \nu_2, \ldots, \nu_q)$ is the $(p + 2)$ vector of parameters.

The loglikelihood function of the parameter ν can be written as

$$l(\nu|r_i, y_i, x_{ij}) = \sum_{i=1}^{n} \big[r_i * \log[\frac{p(r_i = 1)}{1 - p(r_i = 1)}] - \log(1 - p(r_i = 1))\big]. \tag{3.18}$$

Note that choice of variables for the model of r_i is important. Often many variables in this model are not necessarily significant, and more importantly, parameters in the model for r_i are not the primary interest for estimation. Detailed discussion on this can be found in Ibrahim, Lipsitz and Chen (1999) and Ibrahim, Chen and Lipsitz (2001).

Following Ibrahim, Lipsitz and Chen (1999), after incorporating the model for

missingness mechanism ($l(\nu|r_i, y_i, x_{ij})$), the data log likelihood becomes

$$\begin{aligned} l(\psi, \nu|Y, X) = \quad & \sum_{i=1}^{n} \Big[-\log(1+\gamma) + \log[\gamma + f(0; \mu_i, c, \omega)] I_{\{y_i=0\}} \\ & + \log f(y_i; \mu_i, c, \omega) I_{\{y_i>0\}} \Big] \\ & + \sum_{i=1}^{n} \big[r_i * \log[\frac{p(r_i=1)}{1-p(r_i=1)}] - \log(1-p(r_i=1)) \big]. \end{aligned} \tag{3.19}$$

It is to be noted that two parts of this likelihood are separate and their parameters are distinct. This characteristics of the log likelihood facilitates the separate maximization. The rest of the estimation procedure under MNAR remains exactly same as the estimation procedure under MAR.

3.3 Simulation Study

A simulation study was conducted to investigate the properties of the estimates, in terms of bias, variance, mean squared errors (MSE) and coverage probability (CP) of estimates. We use data under four scenarios: (i) data are observed completely, (ii) some responses are missing completely at random (MCAR), (iii) some responses are missing at random (MAR) and (iii) some responses are missing at not at random (MNAR). Two sets of simulations are conducted. The first is without any covariate and the second is with a single covariate.

In the case in which there is no covariate, data are generated from the zero-inflated negative binomial model 3.2 with $\mu = 2$, $c = 0.2$ and $\omega = 0.2$. For the case with one covariate we take $\mu_i = \exp(\sum_{j=1}^{2} X_{ij}\beta_j)$ with $\beta_1 = 1$, $\beta_2 = -1$. Note that β_1 is the intercept parameter, hence $x_{i1} = 1$. The regression variable x_{i2} was generated from

$N(1.5, 0.001)$. We consider 5%, 10% and 25% missing observations in the response variable. For empirical coverage probability we take nominal level $\alpha = 0.05$. For the model of MNAR mechanism, we consider two different covariate structure based on the availability of the covariate in the data: (a) for response only model we consider that missingness of the response may depend only on the response that would have been observed, and (b) for the model having one covariate, we consider that missingness of the response may depend not only on the response that would have been observed but also on the available covariate information. Results with no covariate and where data are observed completely are given in Table 3.1, with no covariate under MCAR, MAR and MNAR are given in Table 3.2, with one covariate and where data are observed completely are given in Table 3.3, and with one covariate under MCAR, MAR and MNAR are given in Table 3.4.

Simulation results in Table 3.1 show that in the usual situation of completely observed data estimate of μ improves (bias, variance and the MSE decrease) as the sample size increases, where as the estimates of c and ω remain reasonably stable. In terms of coverage probability estimates of all three parameters seem some what liberal (empirical coverage is larger than the nominal coverage of 95%).

Results in Table 3.2 show that for MCAR, MAR and MNAR the properties of the estimates of the parameters is similar to that of the completely observed data scenario (Table 3.1) irrespective of the percentage missing. However, for a fixed sample size, as the percentage missing increases all of bias, variance and MSE of μ increase, where as the effect of missingness seem negligible on the estimates of c and ω.

Results based on the completely observed data with covariates (Table 3.3) show that the estimates of β_1, β_2, c, and ω improve as sample size increases. It seems that

presence of covariates have an effect on the estimates of c and ω. As in Table 3.1 and Table 3.2, all the estimates are liberal in terms of coverage probability.

Results for MCAR, MAR and MNAR with covariates (Table 3.4) properties of the estimates of β_1, β_2, c, and ω show similar as the results in Table 3.3.

In general, estimation for the parameters under MNAR shows the closest proximity to the estimation of the parameters under completely observed data. It is to be noted that results under MAR and MNAR are very close though the computational burden for MNAR is substantially huge. It is important to identify the ideal missingness mechanism for the data, which is quite intractable in the case of simulated data.

In simulation results, it is observable that estimates of the over dispersion parameter c become biased with very small variance at the presence of covariate. This behaviour of the over dispersion parameter is also present in the example (in the following section) indicating that the over dispersion property of the response is highly influenced by the covariates.

The parameter values chosen for the simulation are completely arbitrary and it is also expected that the estimation procedure developed here would work for any count data with any covariate.

The above results are for data which come from a zero-inflated negative binomial $NB(\mu, c, \omega)$ distribution. We wish to see whether similar properties of the estimates hold when over-dispersed data are generated from another distribution rather than the $NB(\mu, c)$ distribution. Such a distribution that has been used earlier by others (Lawless, 1987, Paul and Banerjee, 1998) is the log-normal (m, σ^2) mixture of the Poisson distribution with $m = \log(\mu) - \frac{1}{2}\log(c+1)$ and $\sigma^2 = \log(c+1)$, where μ and c are the parameters of the $NB(\mu, c)$. The mean and the variance of this mixture

distribution are μ and $\mu(1+\mu c)$, which are the exact same form compared with the mean and the variance of $NB(\mu, c)$. In the situation in which there are covariates we take $\mu_i = \exp(\sum_{j=1}^{p} X_{ij}\beta_j)$. For more details of generating data from the log-normal mixture of the Poisson distribution see Lawless (1987).

The parameter values used to simulate data from the zero-inflated log-normal mixture of the Poisson distribution were the same as those used to generate data from the zero-inflated negative binomial distribution. We also used the same percentages of missing data as those in the previous case.

Results of the simulation study of the zero-inflated log-normal mixture of the Poisson distributed data are given in Table 3.5, Table 3.6, Table 3.7 and Table 3.8. Fortunately, we arrived at very similar conclusions of the results given in these tables as those of the results in Table 3.1, Table 3.2, Table 3.3 and Table 3.4. This shows, perhaps, that the conclusions will remain similar irrespective of the mechanism in which over-dispersed count data are generated. A point to note is that developing theory based on the zero-inflated log-normal mixture of the Poisson distribution becomes unnecessarily complicated. To avoid that we did the robustness study.

In summary, in the situation in which there is no covariate, the bias, variance and MSE of the estimate of μ decrease as the sample size increases, but increase as the percentage of missing observations increase, whereas the estimates of c and ω remain reasonably stable. For fixed sample size, percentage missingness has an effect only on the estimate of μ. In terms of coverage probability, estimates of all three parameters seem somewhat liberal (empirical coverage is larger than the nominal coverage probability).

Properties of the estimates of the parameters in situations where there are co-

variates remain similar to the situation in which there no covariates except that in the former case the presence of covariates show an effect on the estimation of the parameters c, and ω.

3.4 An Illustrative Example

We now analyze a set of data from a prospective study of dental status of school children from Bohning et al. (1999). The children were all 7 years of age at the beginning of the study. Dental status were measured by the decayed, missing and filled teeth (DMFT) index. Only the eight deciduous molars were considered so the smallest possible value of the DMFT index is 0 and the largest is 8. The prospective study was for a period of two years. The DMFT index was obtained at the beginning of the study and also at the end of the study.

The data also involved 3 categorical covariates: gender having two categories (0 - female, 1 - male), ethnic group having three categories (1 - dark, 2 - white, 3 - black) and school having six categories (1 - oral health education, 2 - all four methods together, 3 - control school (no prevention measure), 4 - enrichment of the school diet with ricebran, 5 - mouthrinse with 0.2% NaF-solution, 6 - oral hygiene).

For the purpose of illustration of our method we deal with the DMFT index data obtained at the beginning of the study (as in Deng and Paul, 2005). The DMFT index data at the beginning of the study are: (index, frequency): (0,172), (1,73), (2,96), (3,80), (4,95), (5,83), (6,85), (7,65), (8,48). We first fit a zero-inflated negative binomial model to the complete data and data with missing observations without covariates. To obtain data with missing observations we randomly deleted a certain

percentage (5%, 10%, 25%) of the observed responses. For estimation under MNAR, we keep simplicity and assumed that missingness may depend only on the missing responses that would have been observed.

The estimates of the mean parameter μ, the over dispersion parameter c and the zero inflation parameter ω based on the zero-inflated negative binomial model, under different percentages of missingness, and their corresponding standard errors are presented in Table 3.9. It is interesting to note that the estimates of the parameters μ, c and ω and the corresponding standard errors remain stable irrespective of the amount of missingness, although, only for MAR and MNAR and for 25% missing values their values are slightly different (slightly larger in case of μ and c). These findings are broadly similar to what was found in the simulation study.

For a little bit more insight we calculated $\hat{E(Y)} = (1-\hat{\omega})\hat{\mu}$ and $\hat{Var(Y)} = (1-\hat{\omega})\hat{\mu}[1+(\hat{c}+\hat{\omega})\hat{\mu}]$ for these data which are given in Table 3.9. It is also interesting to see that these estimates do not vary very much irrespective of the amount of missingness, except under MNAR and 25% missing, where the $\hat{Var(Y)}$ is slightly higher compared to others.

We then fitted a zero-inflated negative binomial model to the complete data and data with missing observations and covariates. Response data with missingness have been obtained exactly the same way as in the situation without covariates. The model fitted was $\mu = \exp(\beta+\beta_{G(M)}I(Gender=1)+\beta_{E(D)}I(Ethnic=1)+\beta_{E(W)}I(Ethnic=2)+\beta_{S(1)}I(School=1)+\beta_{S(2)}I(School=2)+\beta_{S(3)}I(School=3)+\beta_{S(4)}I(School=4)+\beta_{S(5)}I(School=5))$, where β represents the intercept parameter and β_G represents the regression parameter for gender, $\beta_{E(1)}$ and $\beta_{E(2)}$ represent the regression parameters for the ethnic groups 1 and 2, and $\beta_{S(1)}$, $\beta_{S(2)}$, $\beta_{S(3)}$, $\beta_{S(4)}$, and $\beta_{S(5)}$ rep-

resent the regression parameters for school 1, school 2, school 3, school 4, and school 5 respectively.

Estimates of the parameters are given in Table 3.10. In this case the estimates differ (this is expected as it depends on which observations have remained in the final data set). In general, the standard errors of the estimates are larger (in some cases these are much larger, for example, in case of $SE(\hat{\beta_{S(5)}})$) than those under complete data. For MCAR, MAR and MNAR, and 25% missing responses, the standard error is close to twice for missing data in comparison to those for complete data. However, estimates of $E\hat{(Y)}$ do not vary much irrespective of the percentage missing and the missing data mechanism. The same comment applies to $\hat{Var}(Y)$, although for MAR and 25% missing values this is much larger (12.415) than in the other cases (varies between 8.26 to 9.5).

3.5 Discussion

We have an developed estimation procedure for the parameters of a zero inflated negative binomial model in presence of missing observations (responses). We applied a weighted expectation- maximization algorithm (Ibrahim, 1990) for the maximum likelihood estimation of the parameters. Although missing data methodologies have been developed extensively in literature, the current development for the estimation of the parameters of a zero inflated negative binomial model in presence of missing responses is new.

The overall finding of the simulation study is that in the situation in which there is no covariate, the bias, variance and MSE of the estimate of μ decrease as the

sample size increases, but increase as the percentage of missing observations increase, whereas the estimates of c and ω remain reasonably stable. For fixed sample size situations, percentage missingness seems to have an effect only on the estimate of μ. In terms of coverage probability, estimates of all three parameters seem somewhat liberal (empirical coverage is larger than the nominal coverage probability).

Properties of the estimates of the parameters in situations where there are covariates remain similar to the situation in which there no covariates except for the fact that in the former case, the presence of covariates show an effect on the estimation of the parameters c, and ω.

These conclusions remain similar when count data are generated from a log-normal mixture of the Poisson distribution. This possibly shows robustness of the procedure irrespective of the mechanism in which over-dispersed count data are observed.

Table 3.1: *Properties (estimate, bias, variance, mse, coverage probability (cp)) of the estimates of the parameters, data simulated from NB(μ, c, ω), based on 5000 simulation runs*

n	Parameter	$\mu = 2$	$c = 0.2$	$\omega = 0.2$
30	Estimate	2.11	0.17	0.17
	Bias	0.11	-0.03	-0.03
	Variance	0.32	0.29	0.03
	MSE	0.18	0.06	0.01
	CP	0.97	1.00	0.98
50	Estimate	2.06	0.18	0.17
	Bias	0.06	-0.02	-0.03
	Variance	0.20	0.16	0.01
	MSE	0.12	0.06	0.01
	CP	0.96	1.00	0.98
100	Estimate	2.02	0.19	0.16
	Bias	0.02	-0.01	-0.04
	Variance	0.08	0.05	0.01
	MSE	0.06	0.03	0.00
	CP	0.97	0.98	0.99

Table 3.2: *Properties (estimate, bias, variance, mse, coverage probability (cp)) of the estimates of the parameters, data simulated from NB(μ, c, ω), based on 5000 simulation runs*

n		% missing	Missingness Mechanism								
			MCAR	MAR	MNAR	MCAR	MAR	MNAR	MCAR	MAR	MNAR
			$\mu = 2$			$c = 0.2$			$\omega = 0.2$		
30	Estimate	5	2.11	2.18	2.16	0.17	0.13	0.12	0.17	0.18	0.17
		10	2.14	2.20	2.19	0.16	0.12	0.11	0.18	0.18	0.18
		25	2.16	2.40	2.38	0.16	0.06	0.05	0.18	0.20	0.20
	Bias	5	0.11	0.18	0.16	-0.03	-0.07	-0.07	-0.03	-0.02	-0.02
		10	0.14	0.20	0.19	-0.04	-0.08	-0.08	-0.02	-0.02	-0.01
		25	0.16	0.40	0.38	-0.04	-0.14	-0.14	-0.02	0.00	0.00
	Variance	5	0.37	0.44	0.38	0.29	0.32	0.26	0.03	0.03	0.02
		10	0.37	0.55	0.42	0.32	0.46	0.29	0.03	0.03	0.03
		25	0.42	0.56	0.50	0.34	0.33	0.30	0.04	0.02	0.02
	MSE	5	0.19	0.20	0.20	0.05	0.03	0.03	0.01	0.01	0.00
		10	0.21	0.22	0.22	0.07	0.03	0.03	0.01	0.01	0.00
		25	0.25	0.32	0.33	0.05	0.03	0.02	0.01	0.00	0.00
	CP	5	0.98	0.96	0.96	1.00	1.00	1.00	0.98	0.99	0.98
		10	0.97	0.96	0.96	1.00	1.00	1.00	0.98	0.98	0.97
		25	0.97	0.93	0.90	1.00	1.00	1.00	0.98	0.96	0.97
50	Estimate	5	2.05	2.08	2.06	0.19	0.16	0.16	0.17	0.17	0.16
		10	2.06	2.11	2.09	0.19	0.14	0.14	0.17	0.18	0.17
		25	2.08	2.24	2.27	0.18	0.06	0.06	0.17	0.20	0.20
	Bias	5	0.05	0.08	0.06	-0.01	-0.04	-0.03	-0.03	-0.03	-0.03
		10	0.06	0.11	0.09	-0.02	-0.06	-0.05	-0.03	-0.02	-0.03
		25	0.08	0.24	0.27	-0.02	-0.14	-0.13	-0.03	0.00	-0.00
	Variance	5	0.19	0.24	0.23	0.17	0.18	0.17	0.03	0.02	0.01
		10	0.20	0.24	0.25	0.19	0.17	0.18	0.02	0.02	0.01
		25	0.25	0.30	0.32	0.23	0.14	0.16	0.03	0.01	0.00
	MSE	5	0.12	0.12	0.11	0.07	0.03	0.03	0.01	0.01	0.00
		10	0.13	0.13	0.11	0.08	0.03	0.02	0.01	0.00	0.00
		25	0.15	0.15	0.18	0.08	0.03	0.02	0.01	0.00	0.00
	CP	5	0.98	0.97	0.97	1.00	1.00	1.00	0.98	0.99	0.98
		10	0.97	0.96	0.96	1.00	1.00	1.00	0.98	0.98	0.98
		25	0.97	0.94	0.91	1.00	1.00	0.99	0.98	0.98	0.97
100	Estimate	5	2.02	1.99	1.99	0.20	0.21	0.20	0.16	0.16	0.16
		10	2.03	2.02	2.02	0.19	0.19	0.18	0.17	0.17	0.16
		25	2.03	2.19	2.17	0.19	0.08	0.07	0.17	0.20	0.19
	Bias	5	0.02	-0.01	-0.01	0.00	0.01	0.00	-0.04	-0.04	-0.03
		10	0.03	0.02	0.02	-0.01	-0.01	-0.01	-0.03	-0.03	-0.03
		25	0.03	0.19	0.17	-0.01	-0.12	-0.12	-0.03	0.00	-0.00
	Variance	5	0.09	0.10	0.10	0.05	0.08	0.08	0.01	0.01	0.01
		10	0.10	0.12	0.11	0.06	0.09	0.08	0.01	0.01	0.01
		25	0.11	0.13	0.11	0.07	0.06	0.04	0.01	0.01	0.00
	MSE	5	0.07	0.06	0.07	0.03	0.03	0.03	0.01	0.01	0.00
		10	0.07	0.07	0.06	0.03	0.02	0.02	0.00	0.00	0.00
		25	0.08	0.10	0.08	0.03	0.02	0.02	0.00	0.00	0.00
	CP	5	0.96	0.98	0.97	0.98	0.99	0.98	0.99	0.99	0.99
		10	0.97	0.97	0.96	0.99	0.98	0.98	0.99	0.99	0.99
		25	0.97	0.92	0.94	1.00	0.98	0.98	0.99	0.98	0.98

Table 3.3: *Properties (estimate, bias, variance, mse, coverage probability (cp)) of the estimates of the parameters, data simulated from NB(β_1, β_2, c, ω), based on 5000 simulation runs*

n	Parameter	$\beta_1 = 1$	$\beta_2 = -1$	$c = 0.2$	$\omega = 0.2$
30	Estimate	1.20	-1.20	0.03	0.03
	Bias	0.20	-0.20	-0.17	-0.17
	Variance	0.62	0.66	0.80	0.07
	MSE	0.37	0.39	0.09	0.12
	CP	1.00	1.00	1.00	0.98
50	Estimate	1.08	-0.90	0.09	0.22
	Bias	0.08	0.10	-0.11	0.02
	Variance	0.34	0.34	0.64	0.09
	MSE	0.07	0.10	0.78	0.07
	CP	1.00	1.00	1.00	0.94
100	Estimate	1.08	-1.00	0.16	0.19
	Bias	0.08	0.00	-0.04	-0.01
	Variance	0.16	0.17	0.22	0.07
	MSE	0.25	0.21	0.16	0.06
	CP	1.00	1.00	0.99	0.96

Table 3.4: *Properties (estimate, bias, variance, mse, coverage probability (cp)) of the estimates of the parameters, data simulated from NB(β_1, β_2, c, ω), based on 5000 simulation runs*

n		% missing	Missingness Mechanism											
			MCAR	MAR	MNAR	MCAR	MAR	MNAR	MCAR	MAR	MNAR	MCAR	MAR	MNAR
			$\beta_1 = 1$			$\beta_2 = -1$			$c = 0.2$			$\omega = 0.2$		
30	Estimate	5	1.12	1.22	1.16	-0.92	-0.84	-0.81	0.01	0.01	0.01	0.17	0.24	0.25
		10	1.11	1.21	1.06	-0.79	-0.79	-0.79	0.01	0.01	0.01	0.22	0.18	0.23
		25	1.35	0.88	1.29	-1.23	-0.62	-0.82	0.02	0.01	0.01	0.07	0.22	0.31
	Bias	5	0.12	0.22	0.16	0.08	0.16	0.18	-0.19	-0.19	-0.19	-0.03	0.04	0.05
		10	0.11	0.21	0.06	0.21	0.21	0.20	-0.19	-0.19	-0.19	0.02	-0.02	0.03
		25	0.35	-0.12	0.29	-0.23	0.38	0.17	-0.18	-0.19	-0.19	-0.13	0.02	0.11
	Variance	5	0.75	0.67	0.86	0.75	0.66	0.80	0.72	2.41	0.00	0.15	0.02	0.01
		10	0.92	0.66	0.84	0.91	0.65	0.76	0.65	2.78	0.00	0.12	0.02	0.02
		25	0.98	0.74	1.59	0.98	0.74	1.28	0.61	1.03	0.00	0.09	0.02	0.01
	MSE	5	0.07	0.24	0.13	0.10	0.11	0.10	0.04	0.04	0.03	0.05	0.01	0.01
		10	0.11	0.13	0.07	0.14	0.11	0.10	0.04	0.04	0.03	0.03	0.00	0.01
		25	0.29	0.22	0.20	0.39	0.20	0.08	0.04	0.04	0.03	0.33	0.02	0.01
	CP	5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.95	1.00	0.98
		10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.89	1.00	0.95
		25	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.98	0.80	0.84
50	Estimate	5	1.11	1.10	1.06	-0.98	-0.95	-0.88	0.02	0.01	0.01	0.18	0.17	0.19
		10	1.08	1.04	1.10	-1.01	-0.93	-0.92	0.03	0.01	0.01	0.10	0.21	0.20
		25	1.08	1.28	0.94	-1.06	-0.91	-0.75	0.01	0.01	0.01	0.14	0.37	0.27
	Bias	5	0.11	0.10	0.06	0.02	0.05	0.11	-0.18	-0.19	-0.19	-0.02	-0.03	-0.01
		10	0.08	0.04	0.10	-0.01	0.07	0.07	-0.17	-0.19	-0.19	-0.10	0.01	0.000
		25	0.08	0.28	-0.05	-0.06	0.09	0.25	-0.19	-0.19	-0.19	-0.06	0.17	0.07
	Variance	5	0.41	0.40	0.38	0.43	0.40	0.38	0.44	0.93	0.00	0.23	0.02	0.1
		10	0.42	0.42	0.31	0.44	0.42	0.34	0.35	0.80	0.00	0.36	0.02	0.01
		25	0.51	1.12	0.54	0.53	0.91	0.51	0.30	1.20	0.00	0.32	0.01	0.01
	MSE	5	0.15	0.05	0.10	0.11	0.04	0.08	0.04	0.04	0.03	0.03	0.02	0.01
		10	0.11	0.05	0.07	0.13	0.03	0.05	0.07	0.04	0.03	0.14	0.01	0.01
		25	0.40	0.08	0.08	0.33	0.01	0.18	0.04	0.04	0.03	0.30	0.03	0.01
	CP	5	1.00	1.00	0.97	1.00	1.00	0.99	1.00	1.00	0.00	0.95	1.00	0.91
		10	1.00	1.00	0.98	1.00	1.00	1.00	1.00	1.00	0.00	0.95	1.00	0.96
		25	0.98	1.00	1.00	0.97	1.00	1.00	1.00	1.00	0.00	0.97	1.00	0.90
100	Estimate	5	1.05	1.08	1.10	-0.98	-0.97	-0.97	0.11	0.01	0.02	0.15	0.20	0.19
		10	1.00	1.08	1.02	-0.99	-0.96	-0.86	0.18	0.06	0.02	0.13	0.22	0.23
		25	1.07	1.04	0.86	-1.00	-0.92	-0.67	0.08	0.01	0.01	0.15	0.26	0.33
	Bias	5	0.05	0.08	0.10	0.02	0.03	0.02	-0.09	-0.19	-0.18	-0.05	0.00	-0.01
		10	0.00	0.08	0.02	0.01	0.04	0.13	-0.02	-0.14	-0.18	-0.07	0.02	0.03
		25	0.07	0.04	-0.13	0.00	0.08	0.32	-0.12	-0.19	-0.19	-0.05	0.06	0.13
	Variance	5	0.17	0.14	0.15	0.19	0.16	0.16	0.30	0.82	0.00	0.17	0.01	0.01
		10	0.17	0.18	0.16	0.20	0.19	0.18	0.36	1.51	0.00	0.10	0.01	0.01
		25	0.21	0.15	0.15	0.23	0.18	0.19	0.33	1.25	0.00	0.13	0.01	0.01
	MSE	5	0.10	0.12	0.08	0.08	0.09	0.06	0.06	0.04	0.03	0.05	0.01	0.01
		10	0.17	0.11	0.11	0.11	0.10	0.10	0.17	0.04	0.03	0.09	0.01	0.01
		25	0.10	0.03	0.01	0.12	0.04	0.10	0.05	0.04	0.03	0.05	0.01	0.01
	CP	5	1.00	1.00	0.96	0.99	1.00	0.97	0.99	1.00	0.00	0.95	0.93	0.95
		10	1.00	1.00	0.93	1.00	0.96	0.96	1.00	1.00	0.00	0.97	1.00	0.82
		25	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.00	0.96	1.00	1.00

Table 3.5: *Properties (estimate, bias, variance, mse, coverage probability (cp)) of the estimates of the parameters, data simulated from lognormal mixture Poisson (μ, c, ω), based on 5000 simulation runs*

n	Parameter	$\mu = 2$	$c = 0.2$	$\omega = 0.2$
30	Estimate	2.08	0.17	0.17
	Bias	0.08	-0.03	-0.03
	Variance	0.31	0.28	0.05
	MSE	0.17	0.07	0.01
	CP	0.98	1.00	0.99
50	Estimate	2.03	0.18	0.16
	Bias	0.03	-0.02	-0.04
	Variance	0.18	0.13	0.01
	MSE	0.11	0.04	0.01
	CP	0.98	1.00	0.99
100	Estimate	2.00	0.20	0.16
	Bias	0.00	0.00	-0.04
	Variance	0.09	0.06	0.01
	MSE	0.07	0.03	0.01
	CP	0.97	0.98	0.99

Table 3.6: *Properties (estimate, bias, variance, mse, coverage probability (cp)) of the estimates of the parameters, data simulated from lognormal mixture Poisson (μ, c, ω), based on 5000 simulation runs*

n		% missing	Missingness Mechanism								
			MCAR	MAR	MNAR	MCAR	MAR	MNAR	MCAR	MAR	MNAR
			$\mu = 2$			$c = 0.2$			$\omega = 0.2$		
30	Estimate	5	2.09	2.18	2.18	0.16	0.13	0.12	0.17	0.17	0.17
		10	2.08	2.26	2.24	0.17	0.11	0.10	0.17	0.18	0.17
		25	2.15	2.46	2.48	0.16	0.05	0.06	0.17	0.20	0.19
	Bias	5	0.09	0.18	0.18	-0.04	-0.07	-0.07	-0.03	-0.03	-0.02
		10	0.08	0.26	0.24	-0.03	-0.09	-0.09	-0.03	-0.02	-0.02
		25	0.15	0.46	0.48	-0.04	-0.15	-0.13	-0.03	0.00	-0.01
	Variance	5	0.36	0.41	0.40	0.31	0.29	0.28	0.03	0.03	0.02
		10	0.37	0.39	0.40	0.30	0.22	0.26	0.03	0.02	0.02
		25	0.52	0.52	0.53	0.48	0.25	0.28	0.07	0.02	0.02
	MSE	5	0.18	0.20	0.21	0.05	0.03	0.03	0.01	0.01	0.00
		10	0.18	0.25	0.24	0.05	0.03	0.02	0.01	0.01	0.00
		25	0.24	0.41	0.43	0.06	0.03	0.02	0.01	0.00	0.00
	CP	5	0.98	0.96	0.95	1.00	1.00	1.00	0.98	0.98	0.98
		10	0.99	0.96	0.94	1.00	1.00	1.00	0.99	0.98	0.99
		25	0.97	0.88	0.88	1.00	1.00	1.00	0.98	0.96	0.96
50	Estimate	5	2.04	2.08	2.09	0.18	0.16	0.15	0.16	0.17	0.16
		10	2.05	2.14	2.16	0.18	0.12	0.11	0.16	0.18	0.17
		25	2.07	2.40	2.41	0.18	0.07	0.06	0.17	0.20	0.20
	Bias	5	0.04	0.08	0.09	-0.02	-0.04	-0.05	-0.04	-0.03	-0.03
		10	0.05	0.14	0.16	-0.02	-0.08	-0.08	-0.04	-0.02	-0.02
		25	0.07	0.40	0.41	-0.02	-0.13	-0.13	-0.03	0.00	0.00
	Variance	5	0.20	0.23	0.21	0.15	0.16	0.14	0.02	0.02	0.01
		10	0.20	0.21	0.21	0.15	0.12	0.11	0.02	0.01	0.01
		25	0.25	0.30	0.25	0.22	0.14	0.10	0.02	0.01	0.00
	MSE	5	0.11	0.12	0.11	0.04	0.03	0.02	0.01	0.01	0.00
		10	0.12	0.12	0.14	0.04	0.03	0.02	0.01	0.00	0.00
		25	0.14	0.27	0.29	0.06	0.03	0.02	0.01	0.00	0.00
	CP	5	0.98	0.96	0.97	1.00	1.00	0.99	0.99	0.99	0.99
		10	0.97	0.96	0.95	1.00	1.00	1.00	0.99	0.98	0.98
		25	0.98	0.85	0.89	1.00	1.00	1.00	0.98	0.97	0.98
100	Estimate	5	1.99	2.00	2.00	0.20	0.21	0.20	0.16	0.15	0.15
		10	2.00	2.07	2.06	0.20	0.15	0.15	0.16	0.17	0.16
		25	2.00	2.28	2.30	0.20	0.07	0.06	0.16	0.20	0.19
	Bias	5	-0.01	0.00	0.00	0.00	0.01	0.00	-0.04	-0.05	-0.04
		10	0.00	0.07	0.06	0.00	-0.05	-0.04	-0.04	-0.03	-0.03
		25	0.00	0.28	0.30	0.00	-0.13	-0.13	-0.04	0.00	-0.00
	Variance	5	0.09	0.10	0.09	0.05	0.07	0.07	0.01	0.01	0.00
		10	0.09	0.10	0.09	0.06	0.05	0.05	0.01	0.01	0.00
		25	0.11	0.24	0.09	0.08	0.14	0.03	0.01	0.01	0.00
	MSE	5	0.06	0.07	0.06	0.03	0.03	0.02	0.01	0.01	0.00
		10	0.07	0.07	0.06	0.03	0.02	0.02	0.01	0.00	0.00
		25	0.08	0.15	0.14	0.04	0.02	0.02	0.01	0.00	0.00
	CP	5	0.98	0.97	0.97	0.99	0.99	0.99	0.99	0.99	0.99
		10	0.97	0.96	0.96	0.99	0.98	0.99	0.99	1.00	0.99
		25	0.97	0.87	0.88	1.00	0.96	0.98	0.99	0.98	1.00

Table 3.7: *Properties (estimate, bias, variance, mse, coverage probability (cp)) of the estimates of the parameters, data simulated from lognormal mixture Poisson(β_1, β_2, c, ω), based on 5000 simulation runs*

n	Parameter	$\beta_1 = 1$	$\beta_2 = -1$	$c = 0.2$	$\omega = 0.2$
30	Estimate	1.21	-1.11	0.12	0.16
	Bias	0.21	-0.11	-0.08	-0.04
	Variance	1.65	1.55	2.92	0.12
	MSE	0.47	0.47	0.17	0.16
	CP	0.98	0.99	1.00	0.96
50	Estimate	1.09	-1.05	0.18	0.21
	Bias	0.09	-0.05	-0.02	0.01
	Variance	0.35	0.38	1.49	0.12
	MSE	0.45	0.27	0.35	0.12
	CP	0.93	0.96	1.00	0.90
100	Estimate	1.00	-1.00	0.23	0.19
	Bias	0.00	0.00	0.03	-0.01
	Variance	0.16	0.17	1.39	0.16
	MSE	0.21	0.11	0.40	0.12
	CP	0.93	0.96	0.99	0.86

Table 3.8: *Properties (estimate, bias, variance, mse, coverage probability (cp)) of the estimates of the parameters, data simulated from lognormal mixture Poisson(* β_1, β_2, c, ω*), based on 5000 simulation runs*

n		% missing	Missingness Mechanism											
			MCAR	MAR	MNAR	MCAR	MAR	MNAR	MCAR	MAR	MNAR	MCAR	MAR	MNAR
			$\beta_1 = 1$			$\beta_2 = -1$			$c = 0.2$			$\omega = 0.2$		
30	Estimate	5	1.12	1.16	1.05	-0.98	-0.78	-0.75	0.08	0.01	0.01	0.18	0.23	0.24
		10	1.18	1.16	1.01	-1.03	-0.87	-0.77	0.09	0.01	0.00	0.19	0.24	0.23
		25	1.06	1.27	1.07	-0.99	-0.71	-0.84	0.06	0.01	0.00	0.15	0.29	0.25
	Bias	5	0.12	0.16	0.05	0.02	0.22	0.24	-0.12	-0.19	-0.18	-0.02	0.03	0.04
		10	0.18	0.16	0.01	-0.03	0.13	0.22	-0.11	-0.19	-0.19	-0.01	0.04	0.03
		25	0.06	0.27	0.07	0.01	0.29	0.15	-0.14	-0.19	-0.19	-0.05	0.09	0.05
	Variance	5	0.69	0.71	0.77	0.70	0.70	0.73	0.85	3.16	0.00	0.05	0.02	0.02
		10	0.71	0.58	0.87	0.73	0.60	0.81	0.60	1.33	0.00	0.05	0.02	0.02
		25	0.76	1.20	0.69	0.82	1.05	0.72	0.56	3.16	0.00	0.07	0.02	0.02
	MSE	5	0.08	0.44	0.39	0.12	0.30	0.28	0.08	0.04	0.03	0.06	0.01	0.01
		10	0.18	0.10	0.19	0.19	0.06	0.17	0.10	0.04	0.03	0.05	0.01	0.01
		25	0.16	0.23	0.11	0.10	0.11	0.14	0.06	0.04	0.03	0.06	0.02	0.01
	CP	5	1.00	0.98	0.98	1.00	0.98	0.98	1.00	1.00	0.00	0.98	0.97	1.00
		10	1.00	1.00	0.98	1.00	1.00	0.98	1.00	1.00	0.00	0.99	1.00	0.95
		25	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.98	0.90	1.00
50	Estimate	5	1.07	1.06	1.03	-1.01	-0.84	-0.85	0.14	0.01	0.01	0.14	0.22	0.19
		10	1.05	0.98	1.07	-1.03	-0.88	-0.86	0.22	0.01	0.01	0.20	0.19	0.21
		25	1.13	0.84	1.07	-1.04	-0.72	-0.89	0.15	0.04	0.01	0.20	0.29	0.28
	Bias	5	0.07	0.06	0.03	-0.01	0.16	0.14	-0.06	-0.19	-0.19	-0.06	0.02	-0.00
		10	0.05	-0.02	0.07	-0.03	0.12	0.13	0.02	-0.19	-0.19	0.00	-0.01	0.01
		25	0.13	-0.16	0.07	-0.04	0.28	0.10	-0.05	-0.16	-0.19	0.00	0.09	0.08
	Variance	5	0.43	0.35	0.36	0.44	0.37	0.37	0.47	0.76	0.00	0.04	0.02	0.01
		10	0.41	0.31	0.53	0.44	0.35	0.47	0.68	0.53	0.00	0.08	0.02	0.01
		25	0.51	0.40	0.31	0.54	0.43	0.37	0.73	1.47	0.00	0.05	0.01	0.01
	MSE	5	0.18	0.29	0.11	0.16	0.20	0.10	0.16	0.04	0.03	0.12	0.02	0.01
		10	0.20	0.21	0.13	0.16	0.14	0.11	0.34	0.04	0.03	0.13	0.02	0.02
		25	0.24	0.27	0.05	0.24	0.19	0.06	0.22	0.04	0.03	0.10	0.01	0.00
	CP	5	0.97	0.97	0.95	0.98	0.99	1.00	1.00	1.00	0.00	0.87	0.91	0.92
		10	0.98	0.95	0.97	0.99	0.99	1.00	1.00	1.00	0.00	0.90	0.96	0.91
		25	0.98	0.86	1.00	0.99	1.00	1.00	0.99	1.00	0.00	0.89	0.86	1.00
100	Estimate	5	1.01	1.06	1.04	-0.99	-0.93	-0.93	0.23	0.02	0.01	0.19	0.22	0.20
		10	1.03	1.03	1.06	-1.03	-0.94	-0.88	0.20	0.01	0.01	0.14	0.21	0.25
		25	1.06	1.05	0.87	-1.03	-0.72	-0.67	0.23	0.01	0.01	0.19	0.25	0.32
	Bias	5	0.01	0.06	0.04	0.01	0.08	0.06	0.03	-0.18	-0.18	-0.01	0.02	0.00
		10	0.03	0.03	0.06	-0.03	0.06	0.11	0.00	-0.19	-0.18	-0.06	0.01	0.05
		25	0.06	0.05	-0.12	-0.03	0.28	0.32	0.03	-0.19	-0.19	-0.01	0.05	0.12
	Variance	5	0.18	0.15	0.15	0.19	0.17	0.17	0.40	0.39	0.00	0.04	0.01	0.01
		10	0.18	0.15	0.16	0.19	0.17	0.18	0.30	0.36	0.00	0.02	0.01	0.01
		25	0.22	0.21	0.15	0.24	0.23	0.19	0.39	0.18	0.00	0.05	0.01	0.01
	MSE	5	0.27	0.11	0.12	0.16	0.09	0.09	0.36	0.04	0.03	0.09	0.01	0.01
		10	0.17	0.30	0.06	0.12	0.16	0.07	0.27	0.04	0.03	0.08	0.01	0.01
		25	0.25	0.00	0.01	0.17	0.08	0.10	0.47	0.04	0.03	0.10	0.00	0.01
	CP	5	0.94	0.98	0.94	0.96	0.98	0.96	0.97	1.00	0.00	0.84	0.91	0.89
		10	0.94	0.91	1.00	0.96	0.96	1.00	0.99	1.00	0.00	0.82	0.91	0.85
		25	0.94	1.00	1.00	0.97	1.00	1.00	1.00	1.00	0.00	0.87	1.00	1.00

Table 3.9: *Estimates and Standard Errors of the parameters for DMFT index data*

Percentage missingness		$\hat{\mu}$	$SE(\hat{\mu})$	$\hat{c}$	$SE(\hat{c})$	$\hat{\omega}$	$SE(\hat{\omega})$	$\hat{E(y)}$	$\hat{Var(y)}$
Complete data	0%	4.1375	0.0942	0.0530	0.0208	0.1644	0.0107	3.4572	6.5671
MCAR	5%	4.1379	0.0963	0.0533	0.0213	0.1643	0.0109	3.4580	6.5716
	10%	4.1367	0.0987	0.0535	0.0218	0.1643	0.0112	3.4570	6.5717
	25%	4.1381	0.1064	0.0528	0.0235	0.1641	0.0121	3.4590	6.5637
MAR	5%	4.1267	0.0961	0.0640	0.0218	0.1568	0.0105	3.4796	6.6501
	10%	4.1174	0.0984	0.0745	0.0230	0.1499	0.0104	3.5002	6.7341
	25%	4.0576	0.1061	0.1099	0.0278	0.1314	0.0105	3.5122	6.9938
MNAR	5%	4.1174	0.0960	0.0625	0.0217	0.1573	0.0105	3.4697	6.6098
	10%	4.1152	0.0979	0.0732	0.0228	0.1495	0.0104	3.4999	6.7075
	25%	4.0810	0.1058	0.1075	0.0273	0.1330	0.0104	3.5382	7.0109

Table 3.10: *Estimates and Standard Errors of the parameters for DMFT data with covariates*

Percentage missingness		$\hat{\beta}$	$SE(\hat{\beta})$	$\hat{\beta_G}$	$SE(\hat{\beta_G})$	$\hat{\beta_{E(1)}}$	$SE(\hat{\beta_{E(1)}})$	$\hat{\beta_{E(2)}}$	$SE(\hat{\beta_{E(2)}})$
Complete data	0%	0.3863	0.1234	0.0487	0.0517	0.2884	0.0837	0.2407	0.0858
MCAR	5%	0.1555	0.2631	0.1833	0.0764	0.4776	0.1534	0.3837	0.1334
	10%	0.1134	0.1847	0.1746	0.0649	0.4852	0.1154	0.3822	0.1104
	25%	0.1727	0.2215	0.2075	0.0745	0.4712	0.1438	0.3599	0.1379
MAR	5%	0.4765	0.1482	0.0141	0.0563	0.2881	0.0916	0.2174	0.0978
	10%	0.1925	0.1373	0.1110	0.0567	0.3294	0.0866	0.2094	0.0869
	25%	-0.4569	0.2677	0.3002	0.0767	1.0447	0.1710	0.9533	0.1602
MNAR	5%	0.6035	0.1497	0.0145	0.0631	0.0298	0.0960	0.0380	0.0977
	10%	0.5969	0.1249	-0.1991	0.0593	0.3748	0.0883	0.2827	0.0868
	25%	-1.7821	0.2388	-0.1052	0.0792	0.1658	0.1190	0.2140	0.1213

Percentage missingness		$\hat{\beta_{S(1)}}$	$SE(\hat{\beta_{S(1)}})$	$\hat{\beta_{S(2)}}$	$SE(\hat{\beta_{S(2)}})$	$\hat{\beta_{S(3)}}$	$SE(\hat{\beta_{S(3)}})$	$\hat{\beta_{S(4)}}$	$SE(\hat{\beta_{S(4)}})$
Complete data	0%	0.8927	0.1148	0.7948	0.1040	0.9724	0.1126	0.9187	0.1056
MCAR	5%	0.7854	0.2067	0.8824	0.2037	0.8207	0.1915	0.9078	0.1866
	10%	0.8337	0.1472	0.9184	0.1466	0.9102	0.1465	0.9303	0.1388
	25%	0.8396	0.1825	0.8565	0.1701	0.8153	0.1705	0.8727	0.1672
MAR	5%	1.0088	0.1542	0.9331	0.1308	0.5578	0.1100	0.8179	0.1152
	10%	0.9805	0.1233	1.1037	0.1245	0.9938	0.1239	1.0816	0.1204
	25%	0.6305	0.1394	1.0356	0.1790	0.8574	0.1541	0.9786	0.1607
MNAR	5%	0.8136	0.1340	0.9166	0.1309	0.8864	0.1326	0.9065	0.1246
	10%	0.5806	0.1132	0.7807	0.1172	0.7157	0.1157	0.7046	0.1083
	25%	0.8020	0.1587	0.8051	0.1622	0.8687	0.1565	0.6620	0.1549

Percentage missingness		$\hat{\beta_{S(5)}}$	$SE(\hat{\beta_{S(5)}})$	$\hat{c}$	$SE(\hat{c})$	$\hat{\omega}$	$SE(\hat{\omega})$	$\hat{E(y)}$	$\hat{Var(y)}$
Complete data	0%	0.8889	0.1059	0.1327	0.0351	0.1760	0.0159	3.4882	8.0463
MCAR	5%	0.9644	0.2031	0.2279	0.1326	0.1672	0.0300	3.5039	9.3935
	10%	0.9821	0.1502	0.2132	0.0746	0.1653	0.0214	3.4806	9.0386
	25%	0.9360	0.1842	0.2273	0.1081	0.1697	0.0268	3.5225	9.5322
MAR	5%	0.5217	0.1033	0.1993	0.0580	0.2357	0.0199	3.1235	8.6763
	10%	0.9635	0.1123	0.1774	0.0429	0.1608	0.0159	3.4540	8.2625
	25%	1.2090	0.1950	0.3936	0.1169	0.1655	0.0271	3.6227	12.4150
MNAR	5%	0.8727	0.1314	0.2160	0.0696	0.1693	0.0197	3.5215	9.3293
	10%	0.8969	0.1216	0.1893	0.0462	0.1606	0.0159	3.5682	8.8740
	25%	1.0236	0.1451	11.8005	1.6730	-4.1769	0.7366	2.0377	8.1520

Chapter 4

Estimation for Zero Inflated Over dispersed Count Data Model with Missing Covariates

4.1 Introduction

Discrete data in the form of counts often exhibit extra variation that cannot be explained by a simple model, such as the binomial or the Poisson. Also, these data often show more zero counts than what can be predicted by a simple model. For example, Bohning, Dietz, Schlattmann, Mendonca and Kirchner (1999) present a set of data on a prospective study of dental status represented by decayed, missing and filled teeth (DMFT) index of school children from an urban area of Belo Horizonte (Brazil). The data represent decayed, missing and filled teeth (DMFT) index. The DMFT index was observed for 797 children at the beginning and at the end of the

study. The data at the beginning of the study are summarized as (index, frequency): (0,172), (1,73), (2,96), (3,80), (4,95), (5,83), (6,85), (7,65), (8,48). Deng and Paul (2005) fitted the Poisson model, the negative binomial model, the zero-inflated Poison model, and the zero-inflated negative binomial model to these data and showed that the zero-inflated negative binomial model provides the best fit to these data.

As it is well known, analysis of such data, as in the case of normal data (Bohning et al., 1999; Deng and Paul, 2005) may be further complicated when some responses or information on some covariates on some individuals are missing. The purpose of this work is to develop an estimation procedure for the parameters of a zero-inflated negative binomial model when information on some covariates on some individuals are missing.

Using the most popular terminology (Paul and Plackett, 1978; Piegorsch, 1990, Green, 1994; Minami, Cody and Verdesoto, 2007; Mwalili, Lasaffre and Declerck, 2008) the negative binomial random variable with mean parameter μ and dispersion parameter c has the probability mass function

$$f(y;\mu,c) = \quad = \quad \frac{\Gamma(y+c^{-1})}{y!\Gamma(c^{-1})}\left(\frac{c\mu}{1+c\mu}\right)^{y}\left(\frac{1}{1+c\mu}\right)^{c^{-1}}, \tag{4.1}$$

for $y = 0, 1, ..., \mu > 0$. Now, for a typical Y, $Var(Y) = \mu(1+\mu c)$ and $c > -1/\mu$. This is the extended negative binomial distribution of Prentice (1986) which takes account over-dispersion as well as under-dispersion. For $c = 0$, variance of the $NB(\mu, c)$ distribution becomes that of the Poisson(μ) distribution. Further, the limiting distribution of the $NB(\mu, c)$ distribution, as $c \to 0$, is the Poisson(μ).

Using the mass function in equation (4.1) the zero-inflated negative binomial regression model (Deng and Paul, 2005) can be written as

$$f(y_i|x_i;\mu,c,\omega)=\begin{cases}\omega+(1-\omega)(\dfrac{1}{1+c\mu})^{c^{-1}} & \text{if y = 0,}\\ (1-\omega)\dfrac{\Gamma(y+c^{-1})}{y!\Gamma(c^{-1})}\left(\dfrac{c\mu}{1+c\mu}\right)^{y}\left(\dfrac{1}{1+c\mu}\right)^{c^{-1}} & \text{if y > 0}\end{cases} \tag{4.2}$$

with $E(Y)=(1-\omega)\mu$, and $Var(Y)=(1-\omega)\mu[1+(c+\omega)\mu]$, where ω is the zero-inflation parameter. We denote this distribution by $ZINB(\mu,c,\omega)$ distribution.

Extensive work has been done to fit zero-inflated and over-dispersed count data model to real life data (see, for example, Ridout, Demetrio and Hinde, 1998; Hinde and Demetrio, 1998; Li, Lu, Park, Kim, Brinkley and Peterson, 1999; Hall, 2000; Lee, Wang and Yau, 2001; Wang, Lee, Yau and Carrivick, 2003; Lord,Washington and Ivan, 2005; Jiang and Paul,2009; Cameron and Trivedi, 2013).

Also a lot of work has been done to test the presence of zero-inflation and/or over-dispersion (see, for example, Mullahy, 1997; Dean, 1992; Green, 1994; Broek, 1995, Deng and Paul, 2000,2005; Xie, He and Goh, 2001; Paul, Jiang and Balasooriya, 2004; Williamson, Lin, Lyles and Hightower, 2007).

Regression analysis of count data may be further complicated by the existence of missing values either in the response variable and/or in the explanatory variables (covariates). Extensive work has been done on regression analysis of continuous response data with some missing covariates under normality assumption (see, for example, Rubin, 1977; Little and Rubin, 1987, 2002, 2014; Lipsitz and Ibrahim, 1996(a,b); Ibrahim, Chen and Lipsitz, 1999; Ibrahim, Chen, Lipsitz and Herring, 2005; Sinha and Maiti, 2007; Maiti and Pradhan, 2009).

Some work on missing values has also been done on logistic regression analysis

of discrete data (see, for example, Ibrahim, 1990; Lipsitz and Ibrahim, 1996(a,b), Ibrahim, Chen and Lipsitz, 1999, 2001; Ibrahim, Chen, Lipsitz and Herring, 2005; Sinha and Maiti, 2007; Maiti and Pradhan, 2009).

Rubin (1977), and Little and Rubin (1987, 2002, 2014) discuss various missingness mechanisms. If the missingness does not depend on observed data, then the missing data are called missing completely at random (MCAR). If the missing data mechanism depends only on observed data, then the data are missing at random (MAR). The MAR is also known as ignorable missing, that is, in this case, the missing data mechanism is ignored. If the missing data mechanism depends on both observed and unobserved data, that is, failure to observe a value depends on the value that would have been observed, then the data are said to be missing not at random (MNAR) in which case the missingness is nonignorable . For more detailed discussion on missing data mechanism, see Ibrahim et al. (2005).

The purpose of this work is to develop estimation procedure for the parameters of the count data regression model with extra dispersion and zero inflation in presence missing values in the explanatory variables. We specifically use the extended negative binomial model (4.1) as a count data model and missing at random (MAR) scenario as the missing data mechanism. A weighted expectation maximization algorithm (Ibrahim, 1990) is developed for the Maximum likelihood (ML) estimation of the parameters involved. Some simulations are conducted to study the properties of the estimates. Robustness of the procedure is shown when count data follow other over-dispersed models, such as the log-normal mixture of the Poisson distribution. An illustrative example (we use the dental epidemiology data of Bohning et al. (1999) and a discussion leading to some conclusions are given.

The procedure for the estimation of the parameters is developed in Section 2. Results of a simulation study is reported in Section 3. An illustrative example using the dental epidemiology data of Bohning et al. (1999)) is given in Section 4 and a discussion leading to some conclusions is given in Section 5.

4.2 Estimation in Zero-inflated and over-dispersed count data regression model with missing values in the Explanatory Variables

Suppose data for the i^{th} of n subjects are (y_i, x_i), $i = 1, \ldots, n$, which are realizations from $ZINB(\mu, c, \omega)$, where y_i represents the response variable and x_i represents a $p \times 1$ vector of covariates with regression parameter $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$, such that $\mu_i = \exp(\sum_{j=1}^{p} X_{ij}\beta_j)$. Here β_1 is the intercept parameter in which case $X_{i1} = 1$ for all i.

4.2.1 Estimation of the parameters with no missing data

For complete data, the likelihood function is

$$\begin{aligned} L(\beta, c, \omega | y_i) = \prod_{i=1}^{n} \Big[& (\omega + (1-\omega) f(0; \mu_i, c, \omega)) I_{\{y_i = 0\}} \\ & + (1-\omega) f(y_i; \mu_i, c, \omega) I_{\{y_i > 0\}} \Big]. \end{aligned} \tag{4.3}$$

Writing $\gamma = \omega/(1-\omega)$ the log likelihood, apart from a constant, can be written as

$$\begin{aligned} l(\beta, c, \gamma | y_i) = & \sum_{i=1}^{n} \Big[-\log(1+\gamma) + \log[\gamma + f(0; \mu_i, c, \omega)] I_{\{y_i=0\}} \\ & + \log f(y_i; \mu_i, c, \omega) I_{\{y_i>0\}} \Big] \\ = & \sum_{i=1}^{n} \Bigg[-\log(1+\gamma) + \log\big[\gamma + \exp[-c^{-1}\log(1+\mu_i c)]\big] I_{\{y_i=0\}} \\ & + \Big[(y_i \log \mu_i - (y_i + c^{-1}) \log(1+\mu_i c) + \sum_{l=1}^{y_i} [1 + (l-1)c]) \Big] I_{\{y_i>0\}} \Bigg]. \end{aligned} \tag{4.4}$$

The parameters β_j, c and γ can be estimated by directly maximizing the log likelihood function 4.4 or by simultaneously solving the following estimating equations

$$\begin{aligned} \frac{\partial l}{\partial \beta_j} = & \sum_{i=1}^{n} \Bigg[\Big[\frac{-(1+\mu c)^{-1} \exp[(-c^{-1}\log(1+\mu c))]}{\gamma + \exp[(-c^{-1}\log(1+\mu c))]} I_{\{y_i=0\}} \\ & + \big[\frac{y_1}{\mu} - \frac{c(y_1 + c^{-1})}{1+\mu c} \big] I_{\{y_i>0\}} \Big] \frac{\partial \mu_i}{\partial \beta_j} \Bigg] = 0, \end{aligned} \tag{4.5}$$

$$\begin{aligned} \frac{\partial l}{\partial c} = \sum_{i=1}^{n} \Bigg[& \frac{[-\mu c^{-1}(1+\mu c)^{-1} + c^{-2}\log(1+\mu c)] \exp[(-c^{-1}\log(1+\mu c))]}{\gamma + \exp[(-c^{-1}\log(1+\mu c))]} I_{\{y_i=0\}} \\ & + \big[\mu(y_i + c^{-1})(1+\mu c)^{-1} - c^{-2}\log(1+\mu c) + \sum_{l=1}^{y_i} (l-1) \big] I_{\{y_i>0\}} \Bigg] = 0, \end{aligned} \tag{4.6}$$

and

$$\frac{\partial l}{\partial \gamma} = \sum_{i=1}^{n} \Big[-(1+\gamma)^{-1} + \big[\gamma + \exp[(-c^{-1}\log(1+\mu c))]\big]^{-1} I_{\{y_i=0\}} + 0 I_{\{y_i>0\}} \Big] = 0, \tag{4.7}$$

where $\dfrac{\partial \mu_i}{\partial \beta_j} = X_{ij} \exp(\sum_{j=1}^{p} X_{ij}\beta_j)$.

4.2.2 Estimation of the parameters with missing data under MCAR

In the case of MCAR, missingness of the data do not depend on observed data and the subjects having missing observations are deleted before the analysis. For our estimation procedure, the log likelihood function remains the same an given in equation 4.4 with reduced sample size having only complete observations.

4.2.3 Estimation of the parameters with missing data under MAR

As some of the observations in covariates (on some individuals) may be missing we write the covariate x_i as

$$x_i = \begin{cases} x_{o,i} & \text{if } x_i \text{ is observed,} \\ x_{m,i} & \text{if } x_i \text{ is missing.} \end{cases} \tag{4.8}$$

Using this in $f(y_i|x_i;\mu,c,\omega)$ given in equation (4.2), the log-likelihood of y_i $i = 1, ..., n$ or the complete data log-likelihood is

$$\begin{aligned} l(\psi|Y, X_o, X_m) = & \sum_{i=1}^{n} \log(f(y_i|x_i,\psi)) \\ = & \sum_{i=1}^{n} \Big[-\log(1+\gamma) + \log[\gamma + f(0;\mu_i,c,\omega)]I_{\{y_i=0\}} \\ & + \log f(y_i;\mu_i,c,\omega)I_{\{y_i>0\}} \Big], \end{aligned} \tag{4.9}$$

where X_o is the vector of observed values, X_m is the vector of missing values, $\psi = (\beta, c, \gamma)$ and $\mu_i = \exp(\sum_{j=1}^{p} X_{ij}\beta_j)$.

In MAR, the conditional probability of missingness of the data depends on observed data. Parameters of the missingness mechanism are completely separate and distinct from the parameters of the model (4.2). In likelihood based estimation considering MAR, the missingness mechanism can be ignored from the likelihood and missing data that are missing at random are often known as ignorable missing or ignorable non-response, but the subjects having these missing observations cannot be deleted before the analysis (see Little and Rubin, 1987, 2002, 2014 and Ibrahim, Chen, Lipsitz and Herring, 2005 for detailed discussion on this).

In this scenario, our goal is to maximize the following log likelihood (Little and Rubin, 1987, 2002, 2014 p.89) with respect to the parameters ψ

$$l(\psi|Y, X_o) = \sum_{X_m} l(\psi|Y, X_o, X_m). \tag{4.10}$$

For continuous covariates or mixed covariates scenario (Ibrahim, Chen and Lipsitz, 1999) the log likelihood becomes

$$l(\psi|Y, X_o) = \int_{X_m} l(\psi|Y, X_o, X_m) dX_m. \tag{4.11}$$

In the more general case where missing data are not MAR, this likelihood would remain the same but a distribution defining the missing data mechanism needs to be included in the model. For now, this general case is beyond the scope of our research and we concentrate on maximizing $l(\psi|Y, X_o, X_m)$ considering that missing data are MAR.

Direct maximization of $l(\psi|Y, X_o, X_m)$ is not, in general, straight forward. However, the EM algorithm (Dempster, Laird and Rubin, 1977) is a very useful tool for

obtaining maximum likelihood estimates with missing observations.

The EM algorithm uses two iterative steps known as the expectation-step (E-step) and the maximization-step (M-step). Following Little and Rubin (1987, 2002, 2014), the E-step provides the conditional expectation of the log-likelihood $l(\psi|y_i, x_{o,i}, x_{m,i})$ given the observed data $(y_i, x_{o,i})$ and current estimate of the parameters ψ.

Suppose we have a covariate with missing observations and A of the n observations of the covariate are observed and $B = n - A$ observations are missing and s is an arbitrary number of iterations during maximization of the log-likelihood, then the E-step of the EM algorithm for the i^{th} observation of the missing covariate for $(s+1)^{th}$ iteration can be written as

$$\begin{aligned} Q_i(\psi|\psi^{(s)}) &= E[l(\psi|y_i, x_{o,i}, x_{m,i})|y_i, x_{o,i}, \psi^{(s)}] \\ &= \sum_{x_{m,i}} l(\psi|y_i, x_{o,i}, x_{m,i}) P(x_{m,i}|y_i, x_{o,i}, \psi^{(s)}). \end{aligned} \tag{4.12}$$

For continuous covariates or mixed covariates scenario (Ibrahim, Chen, Lipsitz and Herring, 2005) $Q_i(\psi|\psi^{(s)})$ become

$$Q_i(\psi|\psi^{(s)}) = \int_{x_{m,i}} l(\psi|y_i, x_{o,i}, x_{m,i}) P(x_{m,i}|y_i, x_{o,i}, \psi^{(s)}) dx_{m,i}. \tag{4.13}$$

For all the observations, the E-step of EM algorithm for $(s+1)^{th}$ iteration is

$$Q(\psi|\psi^{(s)}) = \sum_{i=1}^{A} l(\psi^{|}y_i, x_i) + \sum_{i=1}^{B} \sum_{x_{m,i}} l(\psi|y_i, x_{o,i}, x_{m,i}) P(x_{m,i}|y_i, x_{o,i}, \psi^{(s)}). \tag{4.14}$$

For all the observations in the case of continuous covariates or mixed covariates cases

(Ibrahim, Chen, Lipsitz and Herring, 2005) $Q(\psi|\psi^{(s)})$ becomes

$$\begin{aligned} Q(\psi|\psi^{(s)}) &= \sum_{i=1}^{A} l(\psi|y_i, x_i) \\ &+ \sum_{i=1}^{B} \int_{x_{m,i}} l(\psi|y_i, x_{o,i}, x_{m,i}) P(x_{m,i}|y_i, x_{o,i}, \psi^{(s)}) dx_{m,i}. \end{aligned} \tag{4.15}$$

Note for the situation in which there is no missing observations in covariates, the EM algorithm requires only maximization of the first term on the right hand side.

Here $P(x_{m,i}|y_i, x_{o,i}, \psi^{(s)})$ is the conditional distribution of the missing covariate given the observed data and the current (s^{th} iteration) estimate of ψ. However, in many situations, $P(x_{m,i}|y_i, x_{o,i}, \psi^{(s)})$ may not always be available. Following Ibrahim, Chen, Lipsitz and Herring, 2005 and Sahu and Roberts, 1999, we can write $P(x_{m,i}|y_i, x_{o,i}, \psi^{(s)}) \propto P(y_i|x_i, \psi^{(s)})P(x_i|\alpha^{(s)})$, where $P(y_i|x_i, \psi^{(s)})$ is the complete data distribution given in (4.2), $P(x_i|\alpha^{(s)})$ is the distribution for the covariates where the missing values exist and both have very elegant forms. For the i^{th} of the B missing observations of the covariate, we take a sample $a_{i1}, a_{i2}, ..., a_{im_i}$ from $P(x_{m,i}|y_i, x_{o,i}, \psi^{(s)})$ using Gibbs sampler (see Casella and George, 1992 for details). Then, following Ibrahim, Chen and Lipsitz (1999) and Ibrahim, Chen, Lipsitz and Herring (2005) $Q(\psi|\psi^{(s)})$ can be written as

$$Q(\psi|\psi^{(s)}) = \sum_{i=1}^{A} l(\psi|y_i, x_i) + \sum_{i=1}^{B} \frac{1}{m_i} \sum_{k=1}^{m_i} l(\psi|y_i, x_{o,i}, a_{ik}). \tag{4.16}$$

In the M-step of the EM algorithm, the $Q(\psi|\psi^{(s)})$ is maximized. Here maximizing $Q(\psi|\psi^{(s)})$ is analogous to maximization of complete data log likelihood with each incomplete covariate being replaced by m_i weighted observations. More details of

EM algorithm by method of weights can be found in Ibrahim, 1990; Lipsitz and Ibrahim, 1996(a,b), Ibrahim, Chen and Lipsitz, 1999, 2001; Ibrahim, Chen, Lipsitz and Herring, 2005; Sinha and Maiti, 2007; Maiti and Pradhan, 2009.

The variance covariance matrix of the estimates of the parameters is calculated by inverting the observed information matrix at convergence (Efron and Hinkley, 1978) which is

$$H_{\psi\psi'} = Q''(\psi|\psi^{(s)}) = \sum_{i=1}^{A} \frac{\partial^2}{\partial\psi\partial\psi'} l(\psi|y_i, x_i) + \sum_{i=1}^{B} \frac{1}{m_i} \sum_{k=1}^{m_i} \frac{\partial^2}{\partial\psi\partial\psi'} l(\psi|y_i, x_{o,i}, a_{ik}) \quad (4.17)$$

Expressions for the elements of H above are given in the Appendix.

4.3 Simulation Study

A simulation study was conducted to investigate the properties of the estimates, in terms of bias, variance, mean squared errors (MSE) and coverage probability (CP) of estimates. We use data under three scenarios: (i) data are observed completely, (ii) some observations in covariates are missing completely at random (MCAR), and (iii) some observations in covariates are missing at random (MAR). Simulations are conducted for continuous as well as discrete covariate.

Responses are generated from the zero-inflated negative binomial model (4.2) with $\mu_i = \exp(\sum_{j=1}^{2} X_{ij}\beta_j)$ where $\beta_1 = 1$, $\beta_2 = -1$, and $c = 0.2$, $\omega = 0.2$. Note that β_1 is the intercept parameter, hence $x_{i1} = 1$. The explanatory variable x_{i2} was generated from $N(1.5, 0.001)$ when covariate is considered to be continuous, and from $Binomial(0.5)$ in case of discrete covariate. We consider 5%, 10% and 25% missing observations in the explanatory variable. For empirical coverage probability we

take nominal level $\alpha = 0.05$. Results with continuous covariate and where data are observed completely are given in Table 4.1, with continuous covariate under MCAR and MAR are given in Table 4.2, with discrete covariate and where data are observed completely are given in Table 4.3, and with discrete covariate under MCAR and MAR are given in Table 4.4.

Simulation results in Table 4.1 show that in the usual situation of completely observed data estimate of β_2 and ω improves (bias, variance and the MSE decrease) as the sample size increases, where as the estimates of c remain reasonably stable and estimate of β_1 increases. In terms of coverage probability, estimates of all three parameters seem somewhat liberal (empirical coverage is larger than the nominal coverage of 95%).

Results in Table 4.2 show that for MCAR and MAR the properties of the estimates of the parameters are similar to that of the completely observed data scenario (Table 4.1) irrespective of the percentage missing. For a fixed sample size, as the percentage missing increases, all of the bias, variance and MSE of β_1, β_2 increase, whereas the effect of missingness seems negligible on the estimates of c and ω. Moreover, for a fixed sample size, irrespective of the percentage missing, estimates of the parameters and their properties (bias, variance and MSE) under MAR remain stable compared to MCAR. Variances of the estimates under MCAR are higher than that of MAR regardless of sample size and percentage missing. Like the completely observed data scenario, all the estimates are liberal in terms of coverage probability.

In Table 4.3, simulation results for completely observed data situation, estimate of β_1, β_2, c and ω improves (bias, variance and the MSE decrease) as the sample size increases. Estimates of β_1 and c are more stable as sample size increases compared to

β_2 and ω. Coverage probability estimates for β_1 and β_2 are very close to the nominal coverage of 95%, whereas coverage probability estimate of c and ω seem somewhat liberal.

Estimates of the parameters β_1, β_2, c and ω and their bias, variance and MSE under MCAR and MAR in Table 4.4 show similar nature as completely observed data scenario in Table 4.3. For a fixed sample size, as the percentage missing increases, all of bias, variance and MSE of all parameters increase though the effect of missingness seems negligible. It is to be noted that variance under MAR is relatively smaller or very close compared to variance of MCAR irrespective of sample size and percentage missing. Coverage probability estimates of the parameters of Table 4.4 show similar characteristics compared to the coverage probabilities in Table 4.3.

Simulation results for the zero-inflated over-dispersed count data regression model under continuous covariate and discrete covariate show the similar nature of the parameters. From the simulation results under discrete covariate, it is to be noted that discrete covariate has an effect on over-dispersion and shows relatively stable results compared to the results under continuous covariate.

In summary, bias, variance and MSE of the estimate of β_1, β_2 decrease as the sample size increases, but increase as the percentage of missing observations increase, whereas the estimates of c and ω remain reasonably stable. For fixed sample size percentage missingness has an effect only on the estimate of β_1, β_2. In terms of coverage probability, estimates of all three parameters seem somewhat liberal (empirical coverage is larger than the nominal coverage probability).

The above results are for data which come from a zero-inflated negative binomial $NB(\beta_1, \beta_2, c, \omega)$ distribution. We wish to see whether similar properties of the esti-

mates hold when over-dispersed data are generated from another distribution rather than the $NB(\beta_1, \beta_2, c)$ distribution. Such a distribution that has been used earlier by others (Lawless, 1987 and Paul and Banergee, 1998) is the log-normal (m, σ^2) mixture of the Poisson distribution with $m = \log(\mu) - \frac{1}{2}\log(c+1)$ and $\sigma^2 = \log(c+1)$, where μ and c are the parameters of the $NB(\mu, c)$. In the situation in which there are covariates we take $\mu_i = \exp(\sum_{j=1}^{p} X_{ij}\beta_j)$. For more details of generating data from the log-normal mixture of the Poisson distribution see Lawless (1987).

The parameter values used to simulate data from the zero-inflated log-normal mixture of the Poisson distribution were the same as those used to generate data from the zero-inflated negative binomial distribution. We also used the same percentages of missing data as those in the previous case.

Results of the simulation study of the zero-inflated log-normal mixture of the Poisson distributed data are given in Table 4.5, Table 4.6, Table 4.7 and Table 4.8 . Fortunately, we arrived at very similar conclusions of the results given in these tables as those of the results in Table 4.1, Table 4.2, Table 4.3 and Table 4.4. This shows, perhaps, that the conclusions will remain similar irrespective of the mechanism in which over-dispersed count data are generated.

4.4 An Illustrative Example

We now analyze a set of data from a prospective study of dental status of school children from Bohning et al. (1999). The children were all 7 years of age at the beginning of the study. Dental status were measured by the decayed, missing and filled teeth (DMFT) index. Only the eight deciduous molars were considered so the

smallest possible value of the DMFT index is 0 and the largest is 8. The prospective study was for a period of two years. The DMFT index was obtained at the beginning of the study and also at the end of the study.

The data also involved 3 categorical covariates: gender having two categories (0 - female, 1 - male), ethnic group having three categories (1 - dark, 2 - white, 3 - black) and school having six categories (1 - oral health education, 2 - all four methods together, 3 - control school (no prevention measure), 4 - enrichment of the school diet with ricebran, 5 - mouthrinse with 0.2% NaF-solution, 6 - oral hygiene).

To illustrate our method, we deal with the DMFT index data obtained at the beginning of the study (as in Deng and Paul, 2005). The DMFT index data at the beginning of the study are: (index, frequency): (0,172), (1,73), (2,96), (3,80), (4,95), (5,83), (6,85), (7,65), (8,48). We then fitted a zero-inflated negative binomial model to the complete data and data with missing observations in covariate. To obtain data with missing observations in covariate we randomly deleted a certain percentage (5%, 10%, 25%) of the observed covariate gender. The model fitted was $\mu = \exp(\beta + \beta_G I(Gender = 1) + \beta_{E(1)} I(Ethnic = 1) + \beta_{E(2)} I(Ethnic = 2) + \beta_{S(1)} I(School = 1) + \beta_{S(2)} I(School = 2) + \beta_{S(3)} I(School = 3) + \beta_{S(4)} I(School = 4) + \beta_{S(5)} I(School = 5))$, where β represents the intercept parameter and β_G represents the regression parameter for gender, $\beta_{E(1)}$ and $\beta_{E(2)}$ represent the regression parameters for the ethnic groups 1 and 2, and $\beta_{S(1)}$, $\beta_{S(2)}$, $\beta_{S(3)}$, $\beta_{S(4)}$, and $\beta_{S(5)}$ represent the regression parameters for school 1, school 2, school 3, school 4, and school 5 respectively.

The estimates of the mean parameter μ, where $\mu_i = \exp(\sum_{j=1}^{p} X_{ij}\beta_j)$ the over dispersion parameter c and the zero inflation parameter ω based on the zero-inflated negative binomial model, under different percentages of missingness, and their corre-

sponding standard errors are presented in Table 4.9. It is to note that the estimates of the parameters μ, c and ω and the corresponding standard errors change with the amount of missingness in the covariate (this is expected as it depends on which observations have remained in the final data set). In general, the standard errors of the estimates are larger than those under complete data. However, estimates of $E(\hat{Y})$ do not vary much irrespective of the percentage missing and the missing data mechanism. The same comment applies to $\hat{Var}(Y)$, although for MCAR and 5% missing values this is much larger (9.4302) than in the other cases (varies between 8.50 to 8.98).

4.5 Discussion

We have developed an estimation procedure for the parameters of a zero inflated negative binomial model in presence of missing observations (covariate). We applied a weighted expectation- maximization algorithm (Ibrahim, 1990) for the maximum likelihood estimation of the parameters. Although missing data methodologies have been developed extensively in the literature, the current development for the estimation of the parameters of a zero inflated negative binomial model in presence of missing covariate is new.

The overall finding of the simulation study is that the bias, variance and MSE of the estimate of the regression parameters decreases as the sample size increases, but increases as the percentage of missing observations increase, whereas the estimates of c and ω remain reasonably stable. For fixed sample size situations percentage missingness seems to have an effect only on the estimate of μ. In terms of coverage probability, estimates of all three parameters seem somewhat liberal (empirical cov-

erage is larger than the nominal coverage probability). It is to be noted that presence of discrete covariates show an effect on the estimation of the parameters c.

These conclusions remain similar when count data are generated from a log-normal mixture of the Poisson distribution. This possibly shows robustness of the procedure irrespective of the mechanism in which over-dispersed count data are observed.

Table 4.1: *Properties (estimate, bias, variance, mse, coverage probability (cp)) of the estimates of the parameters, data simulated from NB(β_1, β_2, c, ω), based on 5000 simulation runs (continuous covariate)*

n	Parameter	$\beta_1 = 1$	$\beta_2 = -1$	$c = 0.2$	$\omega = 0.2$
30	Estimate	1.08	-1.17	0.01	0.04
	Bias	0.08	-0.17	-0.19	-0.16
	Variance	2.16	1.26	0.69	1.10
	MSE	0.02	0.15	0.04	0.05
	CP	1.00	1.00	1.00	1.00
50	Estimate	1.11	-0.90	0.01	0.20
	Bias	0.11	0.10	-0.19	0.00
	Variance	2.80	1.37	0.85	0.40
	MSE	0.22	0.28	0.04	0.07
	CP	0.98	0.98	1.00	0.91
100	Estimate	1.18	-1.02	0.01	0.21
	Bias	0.18	-0.02	-0.19	0.01
	Variance	2.04	0.94	0.32	0.13
	MSE	0.15	0.09	0.04	0.02
	CP	1.00	0.99	1.00	0.95

Table 4.2: *Properties (estimate, bias, variance, mse, coverage probability (cp)) of the estimates of the parameters, data simulated from NB(β_1, β_2, c, ω), based on 5000 simulation runs (continuous covariate)*

n		% missing	Missingness Mechanism							
			MCAR	MAR	MCAR	MAR	MCAR	MAR	MCAR	MAR
			$\beta_1 = 1$		$\beta_2 = -1$		$c = 0.2$		$\omega = 0.2$	
30	Estimate	5	1.12	1.09	-0.92	-0.96	0.01	0.01	0.21	0.21
		10	1.17	1.16	-0.84	-1.04	0.01	0.01	0.22	0.19
		25	1.38	1.14	-1.23	-1.04	0.02	0.01	0.05	0.19
	Bias	5	0.12	0.09	0.08	0.04	-0.19	-0.19	0.01	0.01
		10	0.17	0.16	0.16	-0.04	-0.19	-0.19	0.02	-0.01
		25	0.38	0.14	-0.23	-0.04	-0.18	-0.19	-0.15	-0.02
	Variance	5	7.93	2.15	3.61	0.99	1.02	0.39	0.19	0.07
		10	8.99	2.05	4.05	0.97	6.11	0.36	0.45	0.14
		25	14.51	2.08	6.70	1.10	1.80	0.34	3.69	0.15
	MSE	5	0.06	0.07	0.07	0.07	0.04	0.04	0.02	0.01
		10	0.06	0.13	0.08	0.10	0.04	0.04	0.05	0.02
		25	0.34	0.10	0.40	0.10	0.03	0.04	0.30	0.02
	CP	5	1.00	1.00	1.00	0.99	1.00	1.00	0.97	0.99
		10	1.00	0.99	1.00	0.99	1.00	0.99	0.92	0.97
		25	1.00	1.00	1.00	0.99	1.00	1.00	0.98	0.97
50	Estimate	5	1.10	1.16	-0.98	-1.01	0.02	0.01	0.18	0.21
		10	1.11	1.20	-1.05	-1.06	0.02	0.01	0.08	0.19
		25	1.12	1.17	-1.10	-1.02	0.01	0.01	0.15	0.21
	Bias	5	0.10	0.16	0.02	-0.01	-0.18	-0.19	-0.02	0.01
		10	0.11	0.20	-0.05	-0.06	-0.18	-0.19	-0.12	-0.01
		25	0.12	0.17	-0.10	-0.02	-0.19	-0.19	-0.05	0.01
	Variance	5	12.73	2.14	5.51	1.04	1.28	0.56	0.60	0.11
		10	8.80	2.25	4.39	1.05	1.32	0.31	2.29	0.09
		25	9.58	2.20	4.26	1.03	1.29	0.33	2.71	0.08
	MSE	5	0.09	0.14	0.10	0.11	0.04	0.04	0.03	0.01
		10	0.19	0.16	0.20	0.10	0.04	0.04	0.14	0.02
		25	0.42	0.13	0.34	0.10	0.04	0.04	0.23	0.02
	CP	5	1.00	1.00	1.00	0.99	1.00	1.00	0.96	0.96
		10	1.00	1.00	1.00	0.99	1.00	1.00	0.96	0.98
		25	0.99	1.00	0.98	0.97	1.00	1.00	0.97	0.95
100	Estimate	5	1.09	1.15	-1.03	-1.03	0.13	0.01	0.15	0.18
		10	1.08	1.17	-1.05	-1.00	0.15	0.01	0.13	0.21
		25	1.06	1.17	-0.98	-1.03	0.06	0.01	0.16	0.21
	Bias	5	0.09	0.15	-0.03	-0.03	-0.07	-0.19	-0.05	-0.02
		10	0.08	0.17	-0.05	0.00	-0.05	-0.19	-0.07	0.01
		25	0.06	0.17	0.02	-0.03	-0.14	-0.19	-0.04	0.01
	Variance	5	14.66	2.07	6.64	0.99	1.39	0.43	1.67	0.14
		10	12.00	2.09	5.53	0.97	1.50	0.37	1.51	0.08
		25	10.75	2.08	4.80	1.01	1.10	0.30	1.10	0.08
	MSE	5	0.16	0.12	0.13	0.08	0.15	0.04	0.09	0.02
		10	0.33	0.13	0.26	0.08	0.20	0.04	0.11	0.01
		25	0.09	0.12	0.08	0.08	0.04	0.04	0.05	0.01
	CP	5	1.00	1.00	0.99	0.99	0.98	1.00	0.98	0.96
		10	0.99	1.00	0.98	0.99	0.99	1.00	0.98	0.95
		25	1.00	1.00	1.00	0.99	1.00	1.00	0.95	0.97

Table 4.3: *Properties (estimate, bias, variance, mse, coverage probability (cp)) of the estimates of the parameters, data simulated from NB(β_1, β_2, c, ω), based on 5000 simulation runs (discrete covariate)*

n	Parameter	$\beta_1 = 1$	$\beta_2 = -1$	$c = 0.2$	$\omega = 0.2$
30	Estimate	0.98	-1.04	0.20	0.14
	Bias	-0.02	-0.04	0.00	-0.06
	Variance	0.10	0.19	0.28	0.07
	MSE	0.06	0.21	0.06	0.02
	CP	0.95	0.94	1.00	0.98
50	Estimate	0.97	-1.02	0.21	0.13
	Bias	-0.03	-0.02	0.01	-0.07
	Variance	0.07	0.11	0.20	0.06
	MSE	0.05	0.12	0.07	0.02
	CP	0.96	0.94	1.00	0.98
100	Estimate	0.98	-1.01	0.20	0.15
	Bias	-0.02	-0.01	0.00	-0.05
	Variance	0.03	0.05	0.06	0.01
	MSE	0.03	0.05	0.04	0.01
	CP	0.95	0.94	0.95	0.98

Table 4.4: *Properties (estimate, bias, variance, mse, coverage probability (cp)) of the estimates of the parameters, data simulated from NB(β_1, β_2, c, ω), based on 5000 simulation runs (discrete covariate)*

n		% missing	Missingness Mechanism							
			MCAR	MAR	MCAR	MAR	MCAR	MAR	MCAR	MAR
			$\beta_1 = 1$		$\beta_2 = -1$		$c = 0.2$		$\omega = 0.2$	
30	Estimate	5	0.98	0.97	-1.08	-0.98	0.22	0.23	0.12	0.13
		10	0.98	0.93	-1.08	-0.94	0.22	0.25	0.12	0.11
		25	0.98	0.90	-1.13	-0.76	0.21	0.30	0.12	0.13
	Bias	5	-0.02	-0.03	-0.08	0.02	0.02	0.03	-0.08	-0.07
		10	-0.02	-0.07	-0.08	0.06	0.02	0.05	-0.08	-0.09
		25	-0.02	-0.10	-0.13	0.24	0.01	0.10	-0.08	-0.07
	Variance	5	0.12	0.11	0.24	0.22	0.45	0.47	0.24	0.27
		10	0.14	0.22	0.22	0.20	0.56	0.90	0.35	0.59
		25	0.15	0.16	0.31	0.22	0.59	0.80	0.36	0.44
	MSE	5	0.07	0.07	0.27	0.24	0.10	0.10	0.03	0.03
		10	0.08	0.13	0.25	0.21	0.10	0.15	0.04	0.05
		25	0.09	0.09	0.36	0.28	0.09	0.17	0.04	0.04
	CP	5	0.95	0.95	0.95	0.94	1.00	1.00	0.98	0.98
		10	0.95	0.96	0.95	0.93	1.00	1.00	0.97	0.97
		25	0.94	0.97	0.96	0.90	1.00	1.00	0.98	0.97
50	Estimate	5	0.97	0.95	-1.03	-0.95	0.23	0.24	0.13	0.13
		10	0.97	0.95	-1.05	-0.91	0.22	0.26	0.14	0.13
		25	0.96	0.91	-1.02	-0.74	0.23	0.30	0.12	0.13
	Bias	5	-0.03	-0.05	-0.03	0.05	0.03	0.04	-0.07	-0.07
		10	-0.03	-0.05	-0.05	0.09	0.02	0.06	-0.06	-0.07
		25	-0.04	-0.09	-0.02	0.26	0.03	0.10	-0.08	-0.07
	Variance	5	0.08	0.07	0.12	0.11	0.33	0.26	0.20	0.20
		10	0.06	0.07	0.15	0.12	0.20	0.26	0.15	0.14
		25	0.13	0.08	0.14	0.11	0.51	0.33	0.28	0.28
	MSE	5	0.06	0.06	0.12	0.12	0.10	0.09	0.03	0.03
		10	0.05	0.05	0.16	0.14	0.07	0.10	0.02	0.02
		25	0.07	0.06	0.14	0.19	0.10	0.13	0.03	0.03
	CP	5	0.95	0.96	0.95	0.93	1.00	1.00	0.98	0.98
		10	0.95	0.96	0.96	0.92	1.00	1.00	0.98	0.98
		25	0.95	0.98	0.95	0.85	1.00	1.00	0.98	0.98
100	Estimate	5	0.98	0.96	-1.02	-0.95	0.21	0.24	0.15	0.15
		10	0.98	0.96	-1.01	-0.90	0.21	0.25	0.15	0.15
		25	0.98	0.92	-1.02	-0.73	0.21	0.31	0.15	0.15
	Bias	5	-0.02	-0.04	-0.02	0.05	0.01	0.04	-0.05	-0.05
		10	-0.02	-0.04	-0.01	0.10	0.01	0.05	-0.05	-0.05
		25	-0.02	-0.08	-0.02	0.27	0.01	0.11	-0.05	-0.05
	Variance	5	0.03	0.03	0.06	0.05	0.06	0.09	0.02	0.05
		10	0.03	0.03	0.06	0.05	0.06	0.08	0.01	0.02
		25	0.04	0.03	0.07	0.06	0.10	0.12	0.06	0.06
	MSE	5	0.02	0.03	0.06	0.06	0.04	0.06	0.01	0.01
		10	0.03	0.03	0.07	0.07	0.04	0.06	0.01	0.01
		25	0.03	0.04	0.07	0.13	0.05	0.09	0.02	0.02
	CP	5	0.96	0.95	0.96	0.93	0.96	0.97	0.97	0.98
		10	0.95	0.96	0.95	0.92	0.95	0.97	0.98	0.99
		25	0.96	0.97	0.95	0.75	0.98	0.98	0.98	0.99

Table 4.5: *Properties (estimate, bias, variance, mse, coverage probability (cp)) of the estimates of the parameters, data simulated from Lognormal mixture of Posson(β_1, β_2, c, ω), based on 5000 simulation runs (continuous covariate)*

n	Parameter	$\beta_1 = 1$	$\beta_2 = -1$	$c = 0.2$	$\omega = 0.2$
30	Estimate	1.07	-1.20	0.01	0.02
	Bias	0.07	-0.20	-0.19	-0.18
	Variance	2.00	1.20	0.72	1.18
	MSE	0.02	0.16	0.04	0.06
	CP	1.00	1.00	1.00	1.00
50	Estimate	1.14	-0.91	0.01	0.26
	Bias	0.14	0.09	-0.19	0.06
	Variance	2.81	1.37	0.77	0.20
	MSE	0.25	0.27	0.04	0.06
	CP	0.98	0.98	1.00	0.92
100	Estimate	1.15	-1.02	0.01	0.20
	Bias	0.15	-0.02	-0.19	0.00
	Variance	1.98	0.99	0.39	0.09
	MSE	0.12	0.10	0.04	0.01
	CP	1.00	0.99	1.00	0.96

Table 4.6: *Properties (estimate, bias, variance, mse, coverage probability (cp)) of the estimates of the parameters, data simulated from Lognormal mixture of Posson(β_1, β_2, c, ω), based on 5000 simulation runs (continuous covariate)*

n		% missing	Missingness Mechanism							
			MCAR	MAR	MCAR	MAR	MCAR	MAR	MCAR	MAR
			$\beta_1 = 1$		$\beta_2 = -1$		$c = 0.2$		$\omega = 0.2$	
30	Estimate	5	1.12	1.05	-0.94	-1.11	0.01	0.01	0.18	0.06
		10	1.06	1.09	-0.81	-1.18	0.01	0.01	0.23	0.04
		25	1.37	1.08	-1.28	-1.11	0.02	0.01	0.05	0.07
	Bias	5	0.12	0.05	0.06	-0.11	-0.19	-0.19	-0.02	-0.14
		10	0.06	0.09	0.19	-0.18	-0.19	-0.19	0.03	-0.16
		25	0.37	0.08	-0.28	-0.11	-0.18	-0.19	-0.15	-0.13
	Variance	5	6.95	2.36	3.19	1.32	0.94	0.67	0.29	0.99
		10	7.24	2.18	3.22	1.27	4.22	0.69	0.24	1.10
		25	13.26	2.42	6.34	1.34	1.40	0.65	3.16	0.94
	MSE	5	0.07	0.03	0.08	0.16	0.04	0.04	0.03	0.05
		10	0.09	0.02	0.13	0.15	0.04	0.04	0.02	0.05
		25	0.32	0.03	0.40	0.14	0.03	0.04	0.26	0.05
	CP	5	1.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00
		10	1.00	1.00	1.00	1.00	1.00	1.00	0.90	1.00
		25	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00
50	Estimate	5	1.09	1.09	-0.96	-0.83	0.02	0.01	0.17	0.24
		10	1.05	1.17	-1.02	-0.90	0.02	0.01	0.05	0.24
		25	1.00	1.16	-0.95	-0.89	0.01	0.01	0.19	0.22
	Bias	5	0.09	0.09	0.04	0.17	-0.18	-0.19	-0.03	0.04
		10	0.05	0.17	-0.02	0.10	-0.18	-0.19	-0.15	0.04
		25	0.00	0.16	0.05	0.11	-0.19	-0.19	-0.01	0.02
	Variance	5	12.11	2.90	5.20	1.47	1.27	0.83	0.60	0.13
		10	8.81	3.21	4.38	1.53	1.30	0.72	3.65	0.23
		25	9.73	3.18	4.25	1.58	1.06	0.73	2.53	0.22
	MSE	5	0.07	0.04	0.09	0.09	0.04	0.04	0.04	0.01
		10	0.08	0.23	0.10	0.23	0.03	0.04	0.17	0.03
		25	0.37	0.10	0.27	0.10	0.04	0.04	0.22	0.02
	CP	5	1.00	1.00	1.00	1.00	1.00	1.00	0.96	0.93
		10	1.00	0.98	1.00	0.98	1.00	1.00	0.93	0.92
		25	0.98	1.00	0.99	1.00	0.99	1.00	0.97	0.90
100	Estimate	5	1.04	1.18	-0.99	-1.04	0.11	0.01	0.13	0.20
		10	1.04	1.19	-0.98	-1.05	0.10	0.01	0.14	0.19
		25	1.08	1.13	-1.02	-1.00	0.12	0.01	0.14	0.20
	Bias	5	0.04	0.18	0.01	-0.04	-0.09	-0.19	-0.07	0.00
		10	0.04	0.19	0.02	-0.05	-0.10	-0.19	-0.06	-0.01
		25	0.08	0.13	-0.02	0.00	-0.08	-0.19	-0.06	0.00
	Variance	5	13.85	2.06	6.27	0.99	1.14	0.26	1.46	0.08
		10	11.18	2.03	5.12	1.05	1.29	0.36	1.07	0.13
		25	14.00	1.95	6.35	0.91	1.52	0.25	1.28	0.08
	MSE	5	0.11	0.14	0.09	0.10	0.10	0.04	0.07	0.02
		10	0.16	0.14	0.12	0.09	0.10	0.04	0.06	0.02
		25	0.11	0.12	0.10	0.10	0.10	0.04	0.06	0.02
	CP	5	1.00	1.00	0.99	1.00	0.98	1.00	0.98	0.97
		10	0.99	1.00	1.00	1.00	1.00	1.00	0.97	0.99
		25	1.00	0.99	1.00	0.97	1.00	1.00	0.96	0.97

Table 4.7: *Properties (estimate, bias, variance, mse, coverage probability (cp)) of the estimates of the parameters, data simulated from Lognormal mixture of Poisson(β_1, β_2, c, ω), based on 5000 simulation runs (discrete covariate)*

n	Parameter	$\beta_1 = 1$	$\beta_2 = -1$	$c = 0.2$	$\omega = 0.2$
30	Estimate	0.99	-1.06	0.19	0.14
	Bias	-0.01	-0.06	-0.01	-0.07
	Variance	0.14	0.18	0.33	0.08
	MSE	0.09	0.21	0.06	0.02
	CP	0.94	0.94	1.00	0.97
50	Estimate	0.97	-1.05	0.22	0.13
	Bias	-0.03	-0.05	0.02	-0.07
	Variance	0.06	0.11	0.16	0.04
	MSE	0.04	0.12	0.06	0.02
	CP	0.96	0.95	1.00	0.98
100	Estimate	0.97	-1.01	0.22	0.14
	Bias	-0.03	-0.01	0.02	-0.06
	Variance	0.03	0.05	0.06	0.02
	MSE	0.02	0.06	0.04	0.01
	CP	0.97	0.95	0.95	0.98

Table 4.8: *Properties (estimate, bias, variance, mse, coverage probability (cp)) of the estimates of the parameters, data simulated from Lognormal mixture of Posson(β_1, β_2, c, ω), based on 5000 simulation runs (discrete covariate)*

n		% missing	Missingness Mechanism							
			MCAR	MAR	MCAR	MAR	MCAR	MAR	MCAR	MAR
			$\beta_1 = 1$		$\beta_2 = -1$		$c = 0.2$		$\omega = 0.2$	
30	Estimate	5	0.98	0.95	-1.02	-0.81	0.24	0.22	0.11	0.13
		10	0.96	0.93	-1.11	-0.72	0.23	0.24	0.11	0.13
		25	0.98	0.89	-1.06	-0.51	0.21	0.24	0.11	0.15
	Bias	5	-0.02	-0.05	-0.02	0.19	0.04	0.02	-0.09	-0.07
		10	-0.04	-0.07	-0.11	0.28	0.03	0.04	-0.09	-0.07
		25	-0.02	-0.11	-0.06	0.49	0.01	0.04	-0.09	-0.05
	Variance	5	0.23	0.12	0.23	0.15	0.87	0.36	0.54	0.10
		10	0.13	0.15	0.26	0.15	0.56	0.43	0.38	0.12
		25	0.24	0.19	0.27	0.13	0.89	0.47	0.59	0.11
	MSE	5	0.12	0.08	0.23	0.25	0.12	0.08	0.05	0.03
		10	0.07	0.09	0.27	0.29	0.12	0.09	0.04	0.03
		25	0.13	0.14	0.30	0.42	0.11	0.08	0.05	0.02
	CP	5	0.95	0.96	0.95	0.86	1.00	1.00	0.98	0.97
		10	0.97	0.96	0.96	0.81	1.00	1.00	0.98	0.97
		25	0.95	0.96	0.93	0.67	1.00	1.00	0.99	0.96
50	Estimate	5	0.96	0.93	-1.02	-0.74	0.23	0.25	0.12	0.13
		10	0.97	0.92	-1.03	-0.66	0.23	0.27	0.13	0.13
		25	0.96	0.89	-1.05	-0.47	0.23	0.34	0.11	0.13
	Bias	5	-0.04	-0.07	-0.02	0.26	0.03	0.05	-0.08	-0.07
		10	-0.03	-0.08	-0.03	0.34	0.03	0.07	-0.07	-0.07
		25	-0.04	-0.11	-0.05	0.53	0.03	0.14	-0.09	-0.07
	Variance	5	0.08	0.06	0.12	0.09	0.26	0.20	0.14	0.05
		10	0.08	0.07	0.12	0.08	0.29	0.23	0.19	0.07
		25	0.09	0.10	0.16	0.07	0.38	0.36	0.33	0.10
	MSE	5	0.06	0.05	0.12	0.19	0.08	0.08	0.03	0.02
		10	0.06	0.06	0.12	0.23	0.09	0.09	0.03	0.02
		25	0.06	0.08	0.16	0.37	0.10	0.14	0.04	0.03
	CP	5	0.95	0.96	0.95	0.80	1.00	1.00	0.98	0.98
		10	0.96	0.97	0.95	0.72	1.00	1.00	0.97	0.98
		25	0.96	0.98	0.95	0.44	1.00	1.00	0.98	0.97
100	Estimate	5	0.98	0.93	-1.02	-0.80	0.22	0.27	0.14	0.14
		10	0.97	0.89	-1.01	-0.71	0.22	0.32	0.14	0.13
		25	0.97	0.87	-1.02	-0.42	0.22	0.38	0.14	0.13
	Bias	5	-0.02	-0.07	-0.02	0.20	0.02	0.07	-0.06	-0.06
		10	-0.03	-0.11	-0.01	0.29	0.02	0.12	-0.06	-0.07
		25	-0.03	-0.13	-0.02	0.58	0.02	0.18	-0.06	-0.07
	Variance	5	0.03	0.03	0.06	0.05	0.08	0.09	0.02	0.03
		10	0.03	0.05	0.06	0.04	0.09	0.16	0.05	0.05
		25	0.04	0.05	0.07	0.03	0.10	0.17	0.04	0.05
	MSE	5	0.03	0.04	0.06	0.10	0.05	0.07	0.01	0.01
		10	0.03	0.05	0.06	0.14	0.06	0.11	0.01	0.02
		25	0.03	0.05	0.07	0.38	0.05	0.14	0.01	0.02
	CP	5	0.96	0.97	0.94	0.80	0.97	0.97	0.98	0.98
		10	0.96	0.97	0.94	0.67	0.98	0.99	0.98	0.99
		25	0.96	0.97	0.95	0.15	0.98	0.99	0.98	0.98

Table 4.9: *Estimates and Standard Errors of the parameters for DMFT data with covariates*

Percentage missingness		$\hat{\beta}$	$SE(\hat{\beta})$	$\hat{\beta_G}$	$SE(\hat{\beta_G})$	$\hat{\beta_{E(1)}}$	$SE(\hat{\beta_{E(1)}})$	$\hat{\beta_{E(2)}}$	$SE(\hat{\beta_{E(2)}})$
Complete data	0%	0.3863	0.1234	0.0487	0.0517	0.2884	0.0837	0.2407	0.0858
MCAR	5%	-0.2509	0.2168	0.2202	0.0685	0.8052	0.1382	0.6646	0.1244
	10%	-0.0608	0.1885	0.2257	0.0686	0.7136	0.1259	0.5549	0.1160
	25%	0.3753	0.1538	0.1736	0.0604	0.3478	0.0992	0.3131	0.1010
MAR	5%	0.4447	0.2354	0.0864	0.0654	0.4333	0.1696	0.3555	0.1315
	10%	0.1176	0.1658	0.1673	0.0613	0.5340	0.1071	0.4189	0.1030
	25%	0.3074	0.2274	0.1087	0.0804	0.5196	0.1335	0.4373	0.1213
Percentage missingness		$\hat{\beta_{S(1)}}$	$SE(\hat{\beta_{S(1)}})$	$\hat{\beta_{S(2)}}$	$SE(\hat{\beta_{S(2)}})$	$\hat{\beta_{S(3)}}$	$SE(\hat{\beta_{S(3)}})$	$\hat{\beta_{S(4)}}$	$SE(\hat{\beta_{S(4)}})$
Complete data	0%	0.8927	0.1148	0.7948	0.1040	0.9724	0.1126	0.9187	0.1056
MCAR	5%	0.9654	0.1505	0.9961	0.1540	0.8989	0.1396	1.0009	0.1405
	10%	0.7937	0.1387	0.9279	0.1421	0.7768	0.1249	0.8583	0.1326
	25%	0.7715	0.1429	0.6760	0.1207	0.7764	0.1291	0.7724	0.1279
MAR	5%	0.5412	0.1951	0.6655	0.1620	0.6927	0.2359	0.6524	0.1871
	10%	0.8640	0.1384	0.8504	0.1291	0.8382	0.1246	0.9136	0.1271
	25%	0.5348	0.1503	0.7502	0.1849	0.6413	0.1559	0.6872	0.1470
Percentage missingness		$\hat{\beta_{S(5)}}$	$SE(\hat{\beta_{S(5)}})$	$\hat{c}$	$SE(\hat{c})$	$\hat{\omega}$	$SE(\hat{\omega})$	$\hat{E(y)}$	$\hat{Var(y)}$
Complete data	0%	0.8889	0.1059	0.1327	0.0351	0.1760	0.0159	3.4882	8.0463
MCAR	5%	1.0450	0.1483	0.2551	0.0860	0.1543	0.0227	3.5015	9.4302
	10%	0.9470	0.1359	0.2146	0.0660	0.1567	0.0203	3.4889	8.8842
	25%	0.8257	0.1297	0.1484	0.0492	0.1897	0.0195	3.4618	8.5149
MAR	5%	0.6490	0.1301	0.2130	0.1512	0.1519	0.0315	3.5511	8.9869
	10%	0.9565	0.1328	0.2041	0.0648	0.1430	0.0183	3.6047	8.9114
	25%	0.7537	0.1810	0.2148	0.1287	0.1470	0.0309	3.5762	8.5042

Chapter 5

Estimation for Zero Inflated Over dispersed Count Data Model with Missing Response and Covariates

5.1 Introduction

Discrete data in the form of counts often exhibit extra variation that cannot be explained by a simple model, such as the binomial or the Poisson. Also, these data often show more zero counts than what can be predicted by a simple model. For example, Bohning, Dietz, Schlattmann, Mendonca and Kirchner (1999) present a set of data on a prospective study of dental status represented by decayed, missing and filled teeth (DMFT) index of school children from an urban area of Belo Horizonte (Brazil). The data represent decayed, missing and filled teeth (DMFT) index. The DMFT index was observed for 797 children at the beginning and at the end of the

study. The data at the beginning of the study are summarized as (index, frequency): (0,172), (1,73), (2,96), (3,80), (4,95), (5,83), (6,85), (7,65), (8,48). Deng and Paul (2005) fitted the Poisson model, the negative binomial model, the zero-inflated Poison model, and the zero-inflated negative binomial model to these data and showed that the zero-inflated negative binomial model provide the best fit to these data.

As is well known, analysis of such data, as in the case of normal data (Bohning et al., 1999; Deng and Paul, 2005) may be further complicated when some responses or information on some covariates on some individuals are missing. The purpose of this work is to develop an estimation procedure for the parameters of a zero-inflated negative binomial model when information on some covariates on some individuals are missing.

Using the most popular terminology (Paul and Plackett, 1978; Piegorsch, 1990, Green, 1994; Minami, Cody and Verdesoto, 2007; Mwalili, Lasaffre and Declerck, 2008) the negative binomial random variable with mean parameter μ and dispersion parameter c has the probability mass function

$$f(y;\mu,c) = \quad = \quad \frac{\Gamma(y+c^{-1})}{y!\Gamma(c^{-1})}\left(\frac{c\mu}{1+c\mu}\right)^{y}\left(\frac{1}{1+c\mu}\right)^{c^{-1}}, \tag{5.1}$$

for $y = 0, 1, ..., \mu > 0$. Now, for a typical Y, $Var(Y) = \mu(1+\mu c)$ and $c > -1/\mu$. This is the extended negative binomial distribution of Prentice (1986) which takes account of over-dispersion as well as under-dispersion. For $c = 0$, variance of the $NB(\mu, c)$ distribution becomes that of the Poisson(μ) distribution. Further, the limiting distribution of the $NB(\mu, c)$ distribution, as $c \to 0$, is Poisson(μ).

Using the mass function in equation (5.1) the zero-inflated negative binomial regression model (Deng and Paul, 2005) can be written as

$$f(y_i|x_i;\mu,c,\omega)=\begin{cases}\omega+(1-\omega)(\dfrac{1}{1+c\mu})^{c^{-1}} & \text{if y = 0,}\\ (1-\omega)\dfrac{\Gamma(y+c^{-1})}{y!\Gamma(c^{-1})}\left(\dfrac{c\mu}{1+c\mu}\right)^{y}\left(\dfrac{1}{1+c\mu}\right)^{c^{-1}} & \text{if y > 0}\end{cases} \tag{5.2}$$

with $E(Y)=(1-\omega)\mu$, and $Var(Y)=(1-\omega)\mu[1+(c+\omega)\mu]$, where ω is the zero-inflation parameter. We denote this distribution by $ZINB(\mu,c,\omega)$ distribution.

Extensive work has been done to fit zero-inflated and over-dispersed count data model to real life data (see, for example, Ridout, Demetrio and Hinde, 1998; Hinde and Demetrio, 1998; Li, Lu, Park, Kim, Brinkley and Peterson, 1999; Hall, 2000; Lee, Wang and Yau, 2001; Wang, Lee, Yau and Carrivick, 2003; Lord,Washington and Ivan, 2005; Jiang and Paul,2009; Cameron and Trivedi, 2013).

Also a lot of work has been done to test the presence of zero-inflation and/or over-dispersion (see, for example, Mullahy, 1997; Dean, 1992; Green, 1994; Broek, 1995, Deng and Paul, 2000,2005; Xie, He and Goh, 2001; Paul, Jiang and Balasooriya, 2004; Williamson, Lin, Lyles and Hightower, 2007).

Regression analysis of count data may be further complicated by the existence of missing values either in the response variable and/or in the explanatory variables (co-variates). Extensive work has been done on regression analysis of continuous response data with some missing covariates under normality assumptions (see, for example, Rubin, 1977; Little and Rubin, 1987, 2002, 2014; Lipsitz and Ibrahim, 1996(a,b); Ibrahim, Chen and Lipsitz, 1999; Ibrahim, Chen, Lipsitz and Herring, 2005; Sinha and Maiti, 2007; Maiti and Pradhan, 2009).

Some work on missing values has also been done on logistic regression analysis

of discrete data (see, for example, Ibrahim, 1990; Lipsitz and Ibrahim, 1996(a,b), Ibrahim, Chen and Lipsitz, 1999, 2001; Ibrahim, Chen, Lipsitz and Herring, 2005; Sinha and Maiti, 2007; Maiti and Pradhan, 2009).

Rubin (1977), and Little and Rubin (1987, 2002, 2014) discuss various missingness mechanisms. If the missingness do not depend on observed data, then the missing data are called missing completely at random (MCAR). If the missing data mechanism depends only on observed data, then the data are missing at random (MAR). The MAR is also known as ignorable missing, that is, in this case, the missing data mechanism is ignored. If the missing data mechanism depends on both observed and unobserved data, that is, failure to observe a value depends on the value that would have been observed, then the data are said to be missing not at random (MNAR) in which case the missingness is nonignorable. For more detailed discussion on missing data mechanism see Ibrahim et al. (2005).

The purpose of this paper is to develop an estimation procedure for the parameters of the count data regression model with extra dispersion and zero inflation in presence of missing values in the response as well as explanatory variables. We specifically use the extended negative binomial model (5.1) as a count data model and missing at random (MAR) scenario as the missing data mechanism. A weighted expectation maximization algorithm (Ibrahim, 1990) is developed for the Maximum likelihood (ML) estimation of the parameters involved. Some simulations are conducted to study the properties of the estimates. Robustness of the procedure is shown when count data follow other over-dispersed models, such as the log-normal mixture of the Poisson distribution. An illustrative example (we use the dental epidemiology data of Bohning et al. (1999) and a discussion leading to some conclusions are given.

The procedure for the estimation of the parameters is developed in Section 2. Results of a simulation study is reported in Section 3. An illustrative example using the dental epidemiology data of Bohning et al. (1999)) is given in Section 4 and a discussion leading to some conclusions is given in Section 5.

5.2 Estimation in Zero-inflated and over-dispersed count data regression model with missing values in the Response and Explanatory Variables

Suppose data for the i^{th} of n subjects are (y_i, x_i), $i = 1, \ldots, n$, which are realizations from $ZINB(\mu, c, \omega)$, where y_i represents the response variable and x_i represents a $p \times 1$ vector of covariates with the regression parameter $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$, such that $\mu_i = \exp(\sum_{j=1}^{p} X_{ij}\beta_j)$. Here β_1 is the intercept parameter in which case $X_{i1} = 1$ for all i.

5.2.1 Estimation of the parameters with no missing data

For complete data, the likelihood function is

$$L(\beta, c, \omega | y_i) = \prod_{i=1}^{n} \Big[(\omega + (1-\omega) f(0; \mu_i, c, \omega)) I_{\{y_i = 0\}} + (1-\omega) f(y_i; \mu_i, c, \omega) I_{\{y_i > 0\}} \Big]. \tag{5.3}$$

Writing $\gamma = \omega/(1-\omega)$ the log likelihood, apart from a constant, can be written as

$$\begin{aligned} l(\beta, c, \gamma | y_i) = & \sum_{i=1}^{n} \Big[-\log(1+\gamma) + \log[\gamma + f(0; \mu_i, c, \omega)] I_{\{y_i=0\}} \\ & + \log f(y_i; \mu_i, c, \omega) I_{\{y_i>0\}} \Big] \\ = & \sum_{i=1}^{n} \Bigg[-\log(1+\gamma) + \log\big[\gamma + \exp[-c^{-1}\log(1+\mu_i c)]\big] I_{\{y_i=0\}} \\ & + \Big[(y_i \log \mu_i - (y_i + c^{-1}) \log(1+\mu_i c) + \sum_{l=1}^{y_i} [1 + (l-1)c]) \Big] I_{\{y_i>0\}} \Bigg]. \end{aligned} \tag{5.4}$$

The parameters β_j, c and γ can be estimated by directly maximizing the loglikelihood function 5.4 or by simultaneously solving the following estimating equations

$$\begin{aligned} \frac{\partial l}{\partial \beta_j} = & \sum_{i=1}^{n} \Bigg[\Big[\frac{-(1+\mu c)^{-1} \exp[(-c^{-1}\log(1+\mu c))]}{\gamma + \exp[(-c^{-1}\log(1+\mu c))]} I_{\{y_i=0\}} \\ & + \big[\frac{y_1}{\mu} - \frac{c(y_1 + c^{-1})}{1+\mu c} \big] I_{\{y_i>0\}} \Big] \frac{\partial \mu_i}{\partial \beta_j} \Bigg] = 0, \end{aligned} \tag{5.5}$$

$$\begin{aligned} \frac{\partial l}{\partial c} = \sum_{i=1}^{n} \Bigg[& \frac{[-\mu c^{-1}(1+\mu c)^{-1} + c^{-2}\log(1+\mu c)] \exp[(-c^{-1}\log(1+\mu c))]}{\gamma + \exp[(-c^{-1}\log(1+\mu c))]} I_{\{y_i=0\}} \\ & + \big[\mu(y_i + c^{-1})(1+\mu c)^{-1} - c^{-2}\log(1+\mu c) + \sum_{l=1}^{y_i} (l-1) \big] I_{\{y_i>0\}} \Bigg] = 0, \end{aligned} \tag{5.6}$$

and

$$\frac{\partial l}{\partial \gamma} = \sum_{i=1}^{n} \Big[-(1+\gamma)^{-1} + \big[\gamma + \exp[(-c^{-1}\log(1+\mu c))]\big]^{-1} I_{\{y_i=0\}} + 0 I_{\{y_i>0\}} \Big] = 0, \tag{5.7}$$

where $\dfrac{\partial \mu_i}{\partial \beta_j} = X_{ij} \exp(\sum_{j=1}^{p} X_{ij}\beta_j)$.

5.2.2 Estimation of the parameters with missing data under MCAR

In case of MCAR, missingness of the data do not depend on observed data and the subjects having the missing observations are deleted before the analysis. For estimation procedure the log likelihood function remains the same as given in equation 5.4 with reduced sample size having only complete observations.

5.2.3 Estimation of the parameters with missing data under MAR

As some of the observations in response and covariates (on some individuals) may be missing we write the response y_i as

$$y_i = \begin{cases} y_{o,i} & \text{if } y_i \text{ is observed,} \\ y_{m,i} & \text{if } y_i \text{ is missing.} \end{cases} \tag{5.8}$$

and the covariate x_i as

$$x_i = \begin{cases} x_{o,i} & \text{if } x_i \text{ is observed,} \\ x_{m,i} & \text{if } x_i \text{ is missing.} \end{cases} \tag{5.9}$$

Using this in $f(y_i|x_i;\mu,c,\omega)$ given in equation (5.2), the log-likelihood of y_i $i=1,...,n$ or the complete data log-likelihood is

$$\begin{aligned} l(\psi|Y_o,Y_m,X_o,X_m) = & \sum_{i=1}^{n}\log(f(y_i|x_i,\psi)) \\ = & \sum_{i=1}^{n}\Big[-\log(1+\gamma)+\log[\gamma+f(0;\mu_i,c,\omega)]I_{\{y_i=0\}} \\ & +\log f(y_i;\mu_i,c,\omega)I_{\{y_i>0\}}\Big], \end{aligned} \tag{5.10}$$

where Y_o, X_o are the vector of observed values, Y_m, X_m are the vector of missing values, $\psi=(\beta,c,\gamma)$ and $\mu_i=\exp(\sum_{j=1}^{p}X_{ij}\beta_j)$.

In MAR, conditional probability of missingness of the data depends on observed data. Parameters of the missingness mechanism are completely separate and distinct from the parameters of the model (5.2). In likelihood based estimation considering MAR, missingness mechanism can be ignored from the likelihood and missing data that are missing at random are often known as ignorable missing or ignorable non-response, but the subjects having these missing observations cannot be deleted before the analysis (see Little and Rubin, 1987, 2002, 2014 and Ibrahim, Chen, Lipsitz and Herring, 2005 for detailed discussion on this).

In this scenario, our goal is to maximize the following loglikelihood (Little and Rubin, 1987, 2002, 2014 p.89) with respect to the parameters ψ

$$l(\psi|Y_o,X_o)=\sum_{Y_m}\sum_{X_m}l(\psi|Y_o,Y_m,X_o,X_m). \tag{5.11}$$

For continuous covariates or mixed covariates scenario (Ibrahim, Chen and Lipsitz,

1999) the loglikelihood becomes

$$l(\psi|Y_o, X_o) = \sum_{Y_m} \big[\int_{X_m} l(\psi|Y_o, Y_m, X_o, X_m) dX_m \big]. \tag{5.12}$$

In the more general case where missing data are not MAR, this likelihood would remain the same but a distribution defining the missing data mechanism needs to be included in the model. For now, this general case is beyond the scope of our research and we concentrate on maximizing $l(\psi|Y, X_o, X_m)$ considering that missing data are MAR.

Direct maximization of $l(\psi|Y_o, Y_m, X_o, X_m)$ is not, in general, straight forward. However, the EM algorithm (Dempster, Laird and Rubin, 1977) is a very useful tool for obtaining maximum likelihood estimates with missing observations.

The EM algorithm uses two iterative steps known as the expectation-step (E-step) and the maximization-step (M-step). Following Little and Rubin (1987, 2002, 2014), the E-step provides the conditional expectation of the log-likelihood $l(\psi|y_{o,i}, y_{m,i}, x_{o,i}, x_{m,i})$ given the observed data $(y_{o,i}, x_{o,i})$ and current estimate of the parameters ψ.

Suppose we have response and a covariate with missing observations and A_1 of the n responses and A_2 of the n observations of the covariate are observed and $B_1 = n - A_1$ response and $B_2 = n - A_2$ observations of the covariate are missing, and s is an arbitrary number of iterations during maximization of the log-likelihood, then the E-step of the EM algorithm for the i^{th} missing response and the j^{th} observation of

the missing covariate for $(s+1)^{th}$ iteration can be written as

$$\begin{aligned} Q_{i,j}(\psi|\psi^{(s)}) &= E[l(\psi^{(s)}|y_{o,i}, y_{m,i}, x_{o,i}, x_{m,i})|y_{o,i}, x_{o,i}, \psi^{(s)}] \\ &= \sum_{y_{m,i}} \sum_{x_{m,j}} l(\psi^{(s)}|y_{o,i}, y_{m,i}, x_{o,i}, x_{m,i}) P(y_{m,i}, x_{m,j}|y_{o,i}, x_{o,i}, \psi^{(s)}). \end{aligned} \tag{5.13}$$

For continuous covariates or mixed covariates scenario (Ibrahim, Chen, Lipsitz and Herring, 2005) $Q_i(\psi|\psi^{(s)})$ become

$$\begin{aligned} Q_{i,j}(\psi|\psi^{(s)}) &= \sum_{y_{m,i}} \int_{x_{m,j}} l(\psi^{(s)}|y_{o,i}, y_{m,i}, x_{o,i}, x_{m,i}) \\ &\quad P(y_{m,i}, x_{m,j}|y_{o,i}, x_{o,i}, \psi^{(s)}) dx_{m,i}. \end{aligned} \tag{5.14}$$

For all the observations, the E-step of EM algorithm for $(s+1)^{th}$ iteration is

$$\begin{aligned} Q(\psi|\psi^{(s)}) &= \sum_{i=1}^{A} l(\psi^{(s)}|y_i, x_i) \\ &\quad + \sum_{i=1}^{B} \sum_{y_{m,i}} \sum_{x_{m,j}} l(\psi^{(s)}|y_{o,i}, y_{m,i}, x_{o,i}, x_{m,i}) P(y_{m,i}, x_{m,j}|y_{o,i}, x_{o,i}, \psi^{(s)}) \end{aligned} \tag{5.15}$$

where $A = A_1 \cup A_2$, and $B = B_1 + B_2$.

For all the observations in case continuous covariates or mixed covariates cases (Ibrahim, Chen, Lipsitz and Herring, 2005) $Q(\psi|\psi^{(s)})$ become

$$\begin{aligned} Q(\psi|\psi^{(s)}) &= \sum_{i=1}^{A} l(\psi^{(s)}|y_i, x_i) \\ &\quad + \sum_{i=1}^{B} \sum_{y_{m,i}} \int_{x_{m,j}} l(\psi^{(s)}|y_{o,i}, y_{m,i}, x_{o,i}, x_{m,i}) \\ &\quad P(y_{m,i}, x_{m,j}|y_{o,i}, x_{o,i}, \psi^{(s)}) dx_{m,i}. \end{aligned} \tag{5.16}$$

Note for the situation in which there are no missing observations in covariates the EM algorithm requires only maximization of the first term on the right hand side. Moreover, ignoring the individual with multiple missing cases (i = j; for example, one individual with missing response and more than one missing covariate), the E-step of the EM algorithm leads to the following simple form

$$\begin{aligned} Q(\psi|\psi^{(s)}) &= \sum_{i=1}^{A} l(\psi^{(s)}|y_i, x_i) \\ &+ \sum_{i=1}^{B_1} \sum_{y_{m,i}} l(\psi^{(s)}|y_{o,i}, y_{m,i}, x_{o,i}, x_{m,i}) P(y_{m,i}|y_{o,i}, x_{o,i}, \psi^{(s)}) \\ &+ \sum_{i=1}^{B_2} \sum_{x_{m,j}} l(\psi^{(s)}|y_{o,i}, y_{m,i}, x_{o,i}, x_{m,i}) P(x_{m,j}|y_{o,i}, x_{o,i}, \psi^{(s)}). \end{aligned} \tag{5.17}$$

Here $P(y_{m,i}|y_{o,i}, x_{o,i}, \psi^{(s)})$ and $P(x_{m,j}|y_{o,i}, x_{o,i}, \psi^{(s)})$ are the conditional distributions of the missing response and covariate given the observed data and the current (s^{th} iteration) estimate of ψ respectfully. However, in many situations, $P(y_{m,i}|y_{o,i}, x_{o,i}, \psi^{(s)})$ and $P(x_{m,j}|y_{o,i}, x_{o,i}, \psi^{(s)})$ may not always be available. Following Ibrahim, Chen, Lipsitz and Herring, 2005, and Sahu and Roberts, 1999, we can write $P(y_{m,i}|y_{o,i}, x_{o,i}, \psi^{(s)}) \propto P(y_i|x_i, \psi^{(s)})$ and $P(x_{m,j}|y_i, x_{o,i}, \psi^{(s)}) \propto P(y_i|x_i, \psi^{(s)})P(x_i|\alpha^{(s)})$, where $P(y_i|x_i, \psi^{(s)})$ is the complete data distribution given in (5.2), $P(x_i|\alpha^{(s)})$ is the distribution for the covariates where the missing values exist and both have very elegant forms. For the i^{th} of the B_1 missing responses we take a sample $a_{i1}, a_{i2}, ..., a_{im_i}$ from $P(y_{m,i}|y_{o,i}, x_{o,i}, \psi^{(s)})$, and for the j^{th} of the B_2 missing observations of the covariate we take a sample $a_{j1}, a_{j2}, ..., a_{jm_j}$ from $P(x_{m,i}|y_i, x_{o,i}, \psi^{(s)})$ using Gibbs sampler (see Casella and George, 1992 for details). Then, following Ibrahim, Chen and Lipsitz (1999) and Ibrahim, Chen, Lipsitz and Herring (2005)

$Q(\psi|\psi^{(s)})$ can be written as

$$\begin{aligned} Q(\psi|\psi^{(s)}) &= \sum_{i=1}^{A} l(\psi^{(s)}|y_i, x_i) \\ &+ \sum_{i=1}^{B_1} \frac{1}{m_i} \sum_{k=1}^{m_i} l(\psi^{(s)}|y_{o,i}, x_{o,i}, a_{ik}) \\ &+ \sum_{j=1}^{B_2} \frac{1}{m_j} \sum_{k=1}^{m_j} l(\psi^{(s)}|y_{o,i}, x_{o,i}, a_{jk}). \end{aligned} \tag{5.18}$$

In the M-step of the EM algorithm, the $Q(\psi|\psi^{(s)})$ is maximized. Here maximizing $Q(\psi|\psi^{(s)})$ is analogous to maximization of complete data log likelihood where each incomplete covariate being replaced by m_i weighted observations. More details of EM algorithm by method of weights can be found in Ibrahim, 1990; Lipsitz and Ibrahim, 1996(a,b), Ibrahim, Chen and Lipsitz, 1999, 2001; Ibrahim, Chen, Lipsitz and Herring, 2005; Sinha and Maiti, 2007; Maiti and Pradhan, 2009.

Variance covariance matrix of the estimates of the parameters is calculated by inverting the observed information matrix at convergence (Efron and Hinkley, 1978) which is

$$\begin{aligned} H_{\psi\psi'} = Q''(\psi|\psi^{(s)}) &= \sum_{i=1}^{A} \frac{\partial^2}{\partial\psi\partial\psi'} l(\psi^{(s)}|y_i, x_i) \\ &+ \sum_{i=1}^{B_1} \frac{1}{m_i} \sum_{k=1}^{m_i} \frac{\partial^2}{\partial\psi\partial\psi'} l(\psi^{(s)}|y_{o,i}, x_{o,i}, a_{ik}) \\ &+ \sum_{j=1}^{B_2} \frac{1}{m_j} \sum_{k=1}^{m_j} \frac{\partial^2}{\partial\psi\partial\psi'} l(\psi^{(s)}|y_{o,i}, x_{o,i}, a_{jk}). \end{aligned} \tag{5.19}$$

Expressions for the elements of H above are given in the Appendix.

Chapter 6

Summary and Plan for Future study

6.1 Summary

We have developed an estimation procedure for the parameters of a zero inflated negative binomial model in the presence of missing response and explanatory variables separately. We applied a weighted expectation- maximization algorithm (Ibrahim, 1990) for the maximum likelihood estimation of the parameters. Although missing data methodologies have been developed extensively in the literature, the current development for the estimation of the parameters of a zero inflated negative binomial model in presence of missing data (response and covariates) is new.

The overall finding of the simulation study for the missing response (chapter 3) is that in the situation in which there is no covariate, bias, variance and MSE of the estimate of μ decrease as the sample size increases, but increase as the percentage of

missing observations increase, where as the estimates of c and ω remain reasonably stable. For fixed sample size situations percentage missingness seems to have an effect only on the estimate of μ. In terms of coverage probability, estimates of all three parameters seem somewhat liberal (empirical coverage is larger than the nominal coverage probability). Properties of the estimates of the parameters in situations where there are covariates remain similar to the situation in which there are no covariates except that in the former case, the presence of covariates show an effect on the estimation of the parameters c, and ω.

In our chapter 4, we have developed an estimation procedure for the parameters of a zero inflated negative binomial model in presence of missing explanatory variables. We have considered both discrete as well as continuous variables separately. We have arrived in similar findings compared to the estimation procedure for missing response. Simulation study shows that bias, variance and MSE of the estimate of the regression parameters decrease as the sample size increases, but increase as the percentage of missing observations increase, where as the estimates of c and ω remain reasonably stable. For fixed sample size situations, percentage missingness seems to have an effect only on the estimate of $\beta's$. In terms of coverage probability, estimates of all three parameters seem somewhat liberal (empirical coverage is larger than the nominal coverage probability). It is to be noted that presence of discrete covariates show an effect on the estimation of the parameters c.

These conclusions remain similar when count data are generated from a log-normal mixture of the Poisson distribution. This possibly shows robustness of the procedure irrespective of the mechanism in which over-dispersed count data are observed.

6.2 Plan for Future study: Estimation for the Zero Inflated Over Dispersed Generalized Linear Model(GLM) in the presence of Missing Data

6.2.1 Generalized Linear Model(GLM)

Consider that $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$ are independent observations where each y_i represents response variable and each x_i represents a $p \times 1$ vector of covariates, that is $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $i = 1, 2, \ldots, n$ represent subject. Following Ibrahim (1990), Lipsitz and Ibrahim (1996), Ibrahim and Lipsitz (1996), Ibrahim, Chen and Lipsitz (1999) and Ibrahim, Chen, Lipsitz and Herring (2005), the joint distribution of (y_i, x_i) can be written as the conditional distribution of y_i given x_i and the marginal distribution of x_i. We use the notation $p(y_i|x_i, \psi)$ for the conditional distribution of y_i given x_i and $p(x_i|\alpha)$ for the marginal distribution of x_i. The complete data density of (y_i, x_i) for the subject i can be written as

$$p(y_i, x_i|\psi, \alpha) = p(y_i|x_i, \psi) * p(x_i|\alpha). \tag{6.1}$$

In the conditional distribution $p(y_i|x_i, \psi)$, ψ is the $k \times 1$ vector of parameters. In our model this parameter vector ψ considers regression parameter β through θ, zero inflation parameter $\omega \, or \, \delta$ and over/under dispersion parameter τ, that is $\psi = (\theta, \omega, \tau)$. In the marginal distribution $p(x_i|\alpha)$, α indicates the parameters of covariate distributions.

We consider the natural exponential family distribution for the conditional distribution $p(y_i|x_i, \psi)$. For the following exponential family distribution we consider

parameter θ.

$$p(y_i|x_i,\theta,\varphi) = \exp[\frac{y_i a(\theta_i) - b(\theta_i)}{d_i(\varphi)} + c(y_i,\varphi)] \tag{6.2}$$

where y represents the response variable, $a(\theta_i)$ is the function of mean parameter θ and $d_i(\varphi)$ is the function of scale parameter φ. The parameter θ is used to link the model to the covariates x. Let θ_i be a function of linear predictor η_i, that is $\theta_i = f(\eta_i)$, where f is a monotone differentiable function, known to be the link function and $\eta_i = x_i'\beta$. In η_i, $x_i' = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is the $p \times 1$ vector of covariates and $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ is the $p \times 1$ vector of regression coefficients. If $\theta_i = \eta_i = x_i'\beta$, then the link function f is said to be a canonical link function. We consider $d_i(\varphi) = 1$ throughout our study and hence $p(y_i|x_i,\theta,\varphi)$ would be written as $p(y_i|x_i,\theta)$ or $p(y_i|x_i,\beta)$. The model in (6.2) is known as generalized linear model (GLM). Departure from the generalized linear model can be due to the presence of excessive number of zeros in the data or due to having over/under dispersion features in the data. In generalized linear model, covariates can be discrete or continuous or both. We will describe this feature in next few paragraphs.

6.2.2 Zero Inflated GLM

Generalized linear model with zero inflation has following probability density function (PDF).

$$f_1(y_i|x_i;\omega,\theta) = \begin{cases} \omega + (1-\omega)f(0|x_i;\theta) & \text{if y} = 0 \\ (1-\omega)f(y_i|x_i;\theta) & \text{if y} > 0 \end{cases} \tag{6.3}$$

In $f_1(y_i;\omega,\theta)$, ω is zero inflation/deflation parameter which can take positive as well as negative values. For $\omega < 0$, $f_1(y_i;\omega,\theta)$ indicates zero deflated GLM and for $\omega > 0$, $f_1(y_i;\omega,\theta)$ represents zero inflated GLM. It is very clear that for $\omega = 0$, the zero inflated/deflated GLM $f_1(y_i;\omega,\theta)$ in (6.3) reduces to GLM in (6.2).

6.2.3 Over/Under Dispersed GLM

Suppose that for given θ^*, y has the exponential family model with probability density function

$$p(y_i|x_i,\theta^*) = \exp[y_i\theta_i^* - b(\theta_i^*) + c(y_i)]$$

where θ_i^*'s are continuous independent random variates with finite mean and variance, $E(\theta_i^*) = \theta_i(x_i,\beta)$ and $var(\theta_i^*) = \tau b_i(\theta_i)$ (Dean, 1992). For illustration, considering $\theta^* = \nu\theta$ with $E(\nu) = 1$ and $Var(\nu) = \tau$ implying $E(\theta^*) = \theta, Var(\theta^*) = \tau\theta^2 > 0$. This feature of θ allows extra-exponential variation in the model. Following Cox(1983), Chesher(1984) and Dean (1992) we obtain the mixed model of $p(y_i|x_i,\theta^*)$ by a Taylor Series expansion about θ and taking expectations. The resulting over-dispersed exponential model can be written as,

$$f_2(y_i|x_i;\theta_i,\tau) = f(y_i|x_i;\theta_i)\left\{1 + \sum_{r=2}^{\infty}\frac{\alpha_r}{r!}D_r(y_i|x_i;\theta_i)\right\}$$

where

$$D_r(y_i|x_i,\theta_i) = \left\{\frac{\partial^{(r)}}{\partial\theta_i^{*(r)}}f(y_i|x_i;\theta^*)|_{\theta^*=\theta}\right\}\{f(y_i|x_i;\theta_i)\}^{-1}$$

and $\alpha_r = E(\theta_i^* - \theta_i)^r$. Further, for small τ, we assume that $\alpha_r = o(\tau)$ for $r \geq 3$ then the over dispersed GLM have following pdf

$$f_2(y_i|x_i;\theta_i,\tau) = f(y_i|x_i;\theta_i)\{1 + \frac{\alpha_2}{2!}D_2(y_i|x_i;\theta_i)\},$$

where $\alpha_2 = E(\theta_i^* - \theta_i)^2 = Var(\theta_i^*) = \tau b_i(\theta_i)$. Then $f_2(y_i|x_i;\theta_i,\tau)$ becomes

$$f_2(y_i|x_i;\theta_i,\tau) = f(y_i|x_i;\theta_i)\{1 + \frac{\tau}{2}b_i(\theta_i)D_2(y_i|x_i;\theta_i)\}. \tag{6.4}$$

It is clear that as $\tau \to 0$, the over dispersed GLM $f_2(y_i|x_i;\theta_i,\tau)$ in (6.4) becomes GLM in (6.2).

6.2.4 Zero Inflated Over/Under Dispersed GLM

The generalized linear model in the presence of both zero inflation and over dispersion has the following form

$$f_3(y_i|x_i;\omega,\tau,\theta) = \begin{cases} \omega + (1-\omega)f_2(0|x_i;\tau,\theta) & \text{if y} = 0 \\ (1-\omega)f_2(y_i|x_i;\tau,\theta) & \text{if y} > 0. \end{cases} \tag{6.5}$$

It is obvious that $f_3(y_i|x_i;\omega,\tau,\theta)$ in (6.5) is the general case of (6.4), (6.3) and (6.2).

6.3 Zero Inflated Over/Under Dispersed GLM in the Presence of Missing Covariates

In zero inflated over dispersed GLM $f_3(y_i|x_i;\omega,\tau,\theta)$, we consider that some of the covariates have missing values and missingness mechanism is missing at random (MAR). That is, the conditional probability of missingness may depend on observed data and the unconditional probability of missingness may depend on unobserved data (Ibrahim, Chen, Lipsitz and Herring, 2005). In our research we consider zero inflated over dispersed GLM in the presence of both discrete as well as continuous covariates. Covariate contributions are incorporated in the model (6.1) by using the covariate distribution $p(x_i|\alpha)$. Following Lipsitz and Ibrahim (1996), Ibrahim and Lipsitz (1996), Ibrahim, Chen and Lipsitz (1999) and Ibrahim, Chen, Lipsitz and Herring (2005) a very convenient covariate model can be written as

$$\begin{aligned} p(x_{i1},x_{i2},\ldots,x_{ip}|\alpha) = \; & p(x_{ip}|x_{i1},x_{i2},\ldots,x_{i,p-1},\alpha_p) \\ & \times p(x_{i,p-1}|x_{i1},x_{i2},\ldots,x_{i,p-2},\alpha_{p-1}) \\ & \times p(x_{i,p-2}|x_{i1},x_{i2},\ldots,x_{i,p-3},\alpha_{p-2}) \\ & \times \ldots \times p(x_{i3}|x_{i1},x_{i2},\alpha_3) \times p(x_{i2}|x_{i1},\alpha_2) \times p(x_{i1}|\alpha_1). \end{aligned} \tag{6.6}$$

This covariate model considers the strategy of reducing nuisance parameters in the covariate distribution.

The model we are dealing with is a combination of (6.5) and (6.6), which is

$$\begin{aligned} p(y_i, x_i|\psi, \alpha) &= p(y_i|x_i, \psi) * p(x_i|\alpha) \\ &= f_3(y_i|x_i; \omega, \tau, \theta) * p(x_{i1}, x_{i2}, \ldots, x_{ip}|\alpha) \\ &= \begin{cases} \omega + (1-\omega) f_2(0|x_i; \tau, \theta) & \text{if y} = 0 \\ (1-\omega) f_2(y_i|x_i; \tau, \theta) & \text{if y} > 0. \end{cases} * p(x_{i1}, x_{i2}, \ldots, x_{ip}|\alpha). \end{aligned} \tag{6.7}$$

6.3.1 Zero Inflated Over/Under Dispersed GLM in the Presence of Missing Response

Like the zero inflated over/ under dispersed GLM with missing observations in the covariates, here we consider that the response variable in the zero inflated over dispersed GLM $f_3(y_i|x_i; \omega, \tau, \theta)$ has some missing observations.

6.3.2 EM algorithm by method of weights

In our research we only focus on maximum likelihood estimation (MLE) with the help of expectation maximization (EM) algorithm by the method of weights (Ibrahim, 1990; Lipsitz and Ibrahim ,1996; Ibrahim and Lipsitz ,1996; Ibrahim, Chen and Lipsitz ,1999 and Ibrahim, Chen, Lipsitz and Herring ,2005). EM algorithm can be applied for MLE in two different ways depending on the type of covariate, such as MLE for categorical covariates using weighted EM algorithm and MLE for continuous covariate using Monte Carlo EM algorithm.

We consider zero inflated over/ under dispersed GLM with either categorical covariates or continuous covariates or mixed covariates that are assumed to be MAR.

From (6.1) the complete data density is

$$\begin{aligned} p(y_i, x_i|\psi, \alpha) &= p(y_i|x_i, \psi) * p(x_i|\alpha) \\ &= f_3(y_i|x_i; \omega, \tau, \theta) * p(x_i|\alpha). \end{aligned} \tag{6.8}$$

The log likelihood for the complete data density can be written as

$$\begin{aligned} l(\psi, \alpha|x, y) &= \log(\textstyle\prod_{i=1}^{n} f_3(y_i|x_i; \omega, \tau, \theta) * p(x_i|\alpha)) \\ \Rightarrow l(\Omega|x, y) &= \sum_{i=1}^{n} [\log f_3(y_i|x_i; \omega, \tau, \theta) + \log p(x_i|\alpha)] \\ &= \sum_{i=1}^{n} [l_{y_i|x_i}(\psi) + l_{x_i}(\alpha)], \end{aligned} \tag{6.9}$$

where Ω represents all the parameters of the joint distribution of (y_i, x_i) in equation (6.1), that is $\Omega = (\psi, \alpha) = (\omega, \tau, \theta, \alpha)$. As we have considered that covariates x_i have missing observations, x_i can be written as $x_i = (x_{o,i}, x_{m,i})$, where $x_{o,i}$ represents covariate has been observed for subject i and $x_{m,i}$ represents covariate has been missed for subject i. The dimensions of $x_{o,i}$ and $x_{m,i}$ usually vary along with subject i.

6.3.3 Maximum Likelihood Estimation for Categorical Covariates Using Weighted EM Algorithm

E-step of EM algorithm

The E-step of EM algorithm by the method of weights for the i^{th} observation can be written as

$$\begin{aligned} Q_i(\Omega|\Omega^{(s)}) &= E[l(\Omega|x_i, y_i)|x_{o,i}, y_i, \Omega^{(s)}] \\ &= \sum_{x_{m,i}} P(x_{m,i}|x_{o,i}, y_i, \Omega^{(s)}) l(\Omega|x_i, y_i). \end{aligned} \tag{6.10}$$

For all the observations, the E-step of EM algorithm by the method of weights is

$$\begin{aligned} Q(\Omega|\Omega^{(s)}) &= \sum_{i=1}^{n} Q_i(\Omega|\Omega^{(s)}) \\ &= \sum_{i=1}^{n} \sum_{x_{m,i}} P(x_{m,i}|x_{o,i}, y_i, \Omega^{(s)}) l(\Omega|x_i, y_i) \\ &= \sum_{i=1}^{n} \sum_{x_{m,i}} P(x_{m,i}|x_{o,i}, y_i, \Omega^{(s)}) [l_{y_i|x_i}(\psi) + l_{x_i}(\alpha)] \\ &= \sum_{i=1}^{n} \sum_{x_{m,i}} P(x_{m,i}|x_{o,i}, y_i, \Omega^{(s)}) l_{y_i|x_i}(\psi) + \sum_{i=1}^{n} \sum_{x_{m,i}} P(x_{m,i}|x_{o,i}, y_i, \Omega^{(s)}) l_{x_i}(\alpha) \\ &= \sum_{i=1}^{n} \sum_{x_{m,i}} w_{ij,(s)} l_{y_i|x_i}(\psi) + \sum_{i=1}^{n} \sum_{x_{m,i}} w_{ij,(s)} l_{x_i}(\alpha) \\ &= Q_{[1]}(\psi|\Omega^{(s)}) + Q_{[2]}(\alpha|\Omega^{(s)}), \end{aligned} \tag{6.11}$$

where $w_{ij,(s)} = P(x_{m,i}|x_{o,i}, y_i, \Omega^{(s)})$ and $w_{ij,(s)}$ represents weights for the incomplete observations for subject i and j is indexing for specific covariate pattern for subject i. Weight is the conditional distribution of missing covariates given the observed data and the current estimate of Ω. (s) represents the number of iteration. $w_{ij,(s)} = P(x_{m,i}|x_{o,i}, y_i, \Omega^{(s)})$ can be express as follows by using the Bayes's theorem.

$$
\begin{aligned}
w_{ij,(s)} &= P(x_{m,i}|x_{o,i}, y_i; \Omega^{(s)}) \\
&= \frac{P(y_i, x_{o,i}, x_{m,i(j)}; \Omega^{(s)})}{P(y_i, x_{o,i}; \Omega^{(s)})} \\
&= \frac{P(y_i, x_{o,i}, x_{m,i(j)}; \Omega^{(s)})}{\sum_{x_{m,i}} P(y_i, x_{o,i}, x_{m,i(j)}; \Omega^{(s)})} \\
&= \frac{P(y_i, x_{o,i}, x_{m,i(j)}; \Omega^{(s)})}{\sum_{x_{m,i}} P(y_i, x_i; \Omega^{(s)})} \\
&= \frac{P(y_i, x_{o,i}, x_{m,i(j)}; \Omega^{(s)})}{\sum_{x_{m,i}} P(y_i, x_i|\Omega^{(s)})} \\
&= \frac{P(y_i, x_{o,i}, x_{m,i(j)}; \Omega^{(s)})}{\sum_{x_{m,i}} P(y_i|x_i; \Omega^{(s)}) P(x_i; \Omega^{(s)})} \\
&= \frac{P(y_i|x_{o,i}, x_{m,i(j)}; \Omega^{(s)}) P(x_{o,i}, x_{m,i(j)}; \Omega^{(s)}))}{\sum_{x_{m,i}} P(y_i|x_i; \Omega^{(s)}) P(x_i|\Omega^{(s)})}
\end{aligned}
\tag{6.12}
$$

M-step of EM algorithm

In the M-step of the EM algorithm, $Q(\Omega|\Omega^{(s)})$ is maximized by maximizing $Q_{[1]}(\psi|\Omega^{(s)})$ and $Q_{[2]}(\alpha|\Omega^{(s)})$ separately. Maximizing $Q_{[1]}(\psi|\Omega^{(s)})$ and $Q_{[2]}(\alpha|\Omega^{(s)})$ is analogous to the maximization of complete data log likelihood, where missing observations are replaced by a set of weighted filled in (n_i) observations. n_i is the number of distinct covariate patterns that an observation i could assume given the response y_i and the observed covariate $x_{o,i}$.

6.3.4 Maximum Likelihood Estimation for Continuous or Mixed Covariates Using Monte Carlo EM Algorithm

E-step of Monte Carlo EM algorithm

The E-step of Monte Carlo EM algorithm by the method of weights for the i^{th} observation can be written as following, where summation for the categorical covariate is replaced by integral for the continuous covariates.

$$\begin{aligned} Q_i(\Omega|\Omega^{(s)}) &= E[l(\Omega|x_i, y_i)|x_{o,i}, y_i, \Omega^{(s)}] \\ &= \int_{x_{m,i}} P(x_{m,i}|x_{o,i}, y_i, \Omega^{(s)}) l(\Omega|x_i, y_i) dx_{m,i} \\ &= \int_{x_{m,i}} P(x_{m,i}|x_{o,i}, y_i, \Omega^{(s)}) [l_{y_i|x_i}(\psi) + l_{x_i}(\alpha)] dx_{m,i} \\ &= \int_{x_{m,i}} P(x_{m,i}|x_{o,i}, y_i, \Omega^{(s)}) l_{y_i|x_i}(\psi) dx_{m,i} \\ &+ \int_{x_{m,i}} P(x_{m,i}|x_{o,i}, y_i, \Omega^{(s)}) l_{x_i}(\alpha) dx_{m,i} \end{aligned} \tag{6.13}$$

Following Ibrahim, Chen and Lipsitz (1999) and Ibrahim, Chen, Lipsitz and Herring (2005), $Q_i(\Omega|\Omega^{(s)})$ in (6.13) is evaluated by Monte Carlo EM algorithm of Wei and Tanner (1990) and by using the Gibbs sampler. MLE's are obtained by Gibbs sampler along with the adaptive rejection algorithm provided by Gilks and Wild (1992). The samples are obtained from $P(x_{m,i}|x_{o,i}, y_i, \Omega^{(s)})$, which can be shown as

$$P(x_{m,i}|x_{o,i}, y_i, \Omega^{(s)}) \propto p(y_i|x_i, \psi^s) * p(x_i|\alpha^s). \tag{6.14}$$

It becomes very easy and straightforward to sample from $P(x_{m,i}|x_{o,i}, y_i, \Omega^{(s)})$ by using

the product of $p(y_i|x_i, \psi^s)$ and $p(x_i|\alpha^s)$.

For illustration, we consider that i^{th} observation (subject) has q_i missing observations. For each missing observation a sample $z_{i1}, z_{i2,\ldots,x_{im_i}}$ is obtained from $P(x_{m,i}|x_{o,i}, y_i, \Omega^{(s)})$ by using the Gibbs sampler along with the adaptive rejection algorithm. Let z_{ik} where $k = 1, 2, \ldots, m_i$ be a sample obtained for each missing observations of i^{th} observation (subject). Each z_{ik} is a vector with dimension $q_i \times 1$. This implies that for each missing observations, there are m_i candidate observations are sampled each are weighted by $\frac{1}{m_i}$. Moreover, z_{ik} can also depend on iteration number. The E step of the Monte Carlo EM algorithm of i^{th} observation (subject) for the $(s+1)^{th}$ iteration can be written as

$$Q_i(\Omega|\Omega^{(s)}) = \frac{1}{m_i}\sum_{k=1}^{m_i} l(\Omega|x_{o,i}, z_{i,k}, y_i), \tag{6.15}$$

where $l(\Omega|x_{o,i}, z_{i,k}, y_i)$ is the likelihood of the complete data. In $l(\Omega|x_{o,i}, z_{i,k}, y_i)$ all the missing observations are filled in by m_i candidate observations each with the weight $\frac{1}{m_i}$. This step becomes analogous with the E step of the weighted EM algorithm for the categorical covariates. The E step of Monte Carlo EM algorithm for all the observations can be written as

$$Q(\Omega|\Omega^{(s)}) = \sum_{i=1}^{n} Q_i(\Omega|\Omega^{(s)}) = \sum_{i=1}^{n}\Big[\frac{1}{m_i}\sum_{k=1}^{m_i} l(\Omega|x_{o,i}, z_{i,k}, y_i)\Big]. \tag{6.16}$$

M-step of Monte Carlo EM algorithm

In the M-step of Monte Carlo EM algorithm, $Q(\Omega|\Omega^{(s)})$ is maximized by ordinary complete data maximization by weights.

Appendix A

Appendix

A.1 The Gibbs Sampler

To use the Gibbs sampler, we need to generate each sample point of $a_{i1}, a_{i2}, ..., a_{im_i}$ by using Gibbs sequence. For example, Gibbs sequence for a_{i1} is

$$
\begin{aligned}
&a_{i1}^{(1)} \sim P(y_i^{(0)}|x_i^{(0)}, \psi^{(0)}) \\
&a_{i1}^{(2)} \sim P(y_i^{(0)}|x_i^{(0)}, \psi^{(0)}, a_{i1}^{(1)}) \\
&a_{i1}^{(3)} \sim P(y_i^{(0)}|x_i^{(0)}, \psi^{(0)}, a_{i1}^{(1)}, a_{i1}^{(2)}) \\
&a_{i1}^{(4)} \sim P(y_i^{(0)}|x_i^{(0)}, \psi^{(0)}, a_{i1}^{(1)}, a_{i1}^{(2)}, a_{i1}^{(3)}) \\
&\cdots \\
&a_{i1}^{(k)} \sim P(y_i^{(0)}|x_i^{(0)}, \psi^{(0)}, a_{i1}^{(1)}, a_{i1}^{(2)}, a_{i1}^{(3)}, \cdots, a_{i1}^{(k-1)}).
\end{aligned}
$$

For large K, $a_{i1}^{(k)} = a_{i1}$. According to Sahu and Roberts (1999) $a_{i1}, a_{i2}, ..., a_{im_i}$ can be considered as a block and can obtained from $P(y_i^{(0)}|x_i^{(0)}, \psi^{(0)})$. In this scenario, for each missing response, samples are considered as a block. For example if there are 5

missing response, then there are 5 blocks. Sahu and Roberts (1999) also mentioned that most practical cases, missing observations are independent of parameters and considered as a single block. In this case, 5 missing observations can be treated as a single block. In our model, missing responses are independent of parameters and hence we follow Sahu and Roberts (1999), for Gibbs sampling. We stop the sequence and obtain the required sample for which the absolute deviation of parameters between two consecutive steps becomes minimal. Extensive explanation of Gibbs sampler is available in Casella and George (1992) and Sahu and Roberts (1999).

A.2 Elements of the observed information matrix

From equation (3.13 or 4.16) we have

$$Q(\psi|\psi^{(s)}) = \sum_{i=1}^{A} l(\psi^{(s)}|y_i, x_i) + \sum_{i=1}^{B} \frac{1}{m_i} \sum_{k=1}^{m_i} l(\psi^{(s)}|y_i, x_{o,i}, a_{ik}). \tag{A.1}$$

Maximizing $Q(\psi|\psi^{(s)})$ is analogous to maximization of complete data log likelihood, $l(\beta, c, \gamma|y_i)$ in (3.4, 4.4) with each incomplete response being replaced by m_i weighted observations. The elements of the observed information matrix are as given below.

$$\begin{aligned}\frac{\partial^2 l}{\partial \beta_j^2} = \sum_{i=1}^{n} \Bigg[& \frac{\frac{1}{(1+\mu c)^2}\exp(-c^{-1}\log(1+\mu c))(\gamma + c(\gamma + \exp(-c^{-1}\log(1+\mu c))))}{[\gamma + \exp(-c^{-1}\log(1+\mu c))]^2} \\ & [\tfrac{\partial \mu_i}{\partial \beta_j}]^2 I_{\{y_i=0\}} \\ & + \Big[\frac{-y_i}{\mu^2} + \frac{c^2(y_i + c^{-1})}{(1+\mu c)^2}\Big][\frac{\partial \mu_i}{\partial \beta_j}]^2 I_{\{y_i>0\}} \\ & + \Big[\frac{-(1+\mu c)^{-1}\exp[(-c^{-1}\log(1+\mu c))]}{\gamma + \exp[(-c^{-1}\log(1+\mu c))]} I_{\{y_i=0\}} \\ & + \Big[\frac{y_1}{\mu} - \frac{c(y_1 + c^{-1})}{1+\mu c}\Big] I_{\{y_i>0\}}\Big] \frac{\partial^2 \mu_i}{\partial \beta_j^2}\Bigg]\end{aligned} \tag{A.2}$$

$$\begin{aligned}\frac{\partial^2 l}{\partial \beta_j \beta_j'} = \sum_{i=1}^{n} \Bigg[& \frac{\frac{1}{(1+\mu c)^2}\exp(-c^{-1}\log(1+\mu c))(\gamma + c(\gamma + \exp(-c^{-1}\log(1+\mu c))))}{[\gamma + \exp(-c^{-1}\log(1+\mu c))]^2} \\ & [\tfrac{\partial \mu_i}{\partial \beta_j}\tfrac{\partial \mu_i}{\partial \beta_j'}] I_{\{y_i=0\}} \\ & + \Big[\frac{-y_i}{\mu^2} + \frac{c^2(y_i + c^{-1})}{(1+\mu c)^2}\Big][\frac{\partial \mu_i}{\partial \beta_j}\frac{\partial \mu_i}{\partial \beta_j'}] I_{\{y_i>0\}} \\ & + \Big[\frac{-(1+\mu c)^{-1}\exp[(-c^{-1}\log(1+\mu c))]}{\gamma + \exp[(-c^{-1}\log(1+\mu c))]} I_{\{y_i=0\}} \\ & + \Big[\frac{y_1}{\mu} - \frac{c(y_1 + c^{-1})}{1+\mu c}\Big] I_{\{y_i>0\}}\Big] \frac{\partial^2 \mu_i}{\partial \beta_j \beta_j'}\Bigg]\end{aligned} \tag{A.3}$$

$$\begin{aligned}\frac{\partial^2 l}{\partial \beta_j \partial c} = \sum_{i=1}^{n} \Bigg[& (\frac{1}{1+\mu c})\exp(-c^{-1}\log(1+\mu c))\Big[[\gamma + \exp(-c^{-1}\log(1+\mu c))] \\ & [c^{-1}(\tfrac{\mu}{1+\mu c}) - c^{-2}\log(1+\mu c) + \tfrac{\mu}{1+\mu c}] \\ & + [\exp(-c^{-1}\log(1+\mu c))][-c^{-1}(\tfrac{\mu}{1+\mu c}) + c^{-2}\log(1+\mu c)]\Big] \\ & \Big[[\gamma + \exp(-c^{-1}\log(1+\mu c))]^2\Big]^{-1}[\frac{\partial \mu_i}{\partial \beta_j}] I_{\{y_i=0\}} \\ & + \frac{-(1+\mu c)[c(y_i - c^{-2}) + (y_i + c^{-1})] - \mu c(y_i + c^{-1})}{(1+\mu c)^2}[\frac{\partial \mu_i}{\partial \beta_j}] I_{\{y_i>0\}}\Bigg]\end{aligned} \tag{A.4}$$

$$\frac{\partial^2 l}{\partial \beta_j \partial \gamma} = \sum_{i=1}^{n} \left[\frac{(\frac{1}{1+\mu c}) \exp(-c^{-1}\log(1+\mu c))}{[\gamma + \exp(-c^{-1}\log(1+\mu c))]^2} [\frac{\partial \mu_i}{\partial \beta_j}] I_{\{y_i=0\}} + 0 I_{\{y_i>0\}} \right] \tag{A.5}$$

$$\begin{aligned} \frac{\partial^2 l}{\partial c^2} = \sum_{i=1}^{n} \Bigg[& \exp(-c^{-1}\log(1+\mu c)) \Big[[\gamma + \exp(-c^{-1}\log(1+\mu c))] \\ & \Big[-c^{-1} \tfrac{-\mu^2}{(1+\mu c)^2} + 2c^{-2} \tfrac{\mu}{1+\mu c} + (-2)c^{-3}\log(1+\mu c) \\ & + [c^{-1}(\tfrac{\mu}{1+\mu c}) - c^{-2}\log(1+\mu c)]^2 \Big] \\ & - \exp(-c^{-1}\log(1+\mu c)) [c^{-1}(\tfrac{\mu}{1+\mu c}) - c^{-2}\log(1+\mu c)]^2 \Big] \\ & \Big[[\gamma + \exp(-c^{-1}\log(1+\mu c))]^2 \Big]^{-1} I_{\{y_i=0\}} \\ & + \Big[(y_i + c^{-1}) \tfrac{-\mu^2}{(1+\mu c)^2} + (-2)c^{-2} \tfrac{\mu}{1+\mu c} + 2c^{-3}\log(1+\mu c) \Big] I_{\{y_i>0\}} \Bigg] \end{aligned} \tag{A.6}$$

$$\begin{aligned} \frac{\partial^2 l}{\partial c \partial \gamma} = \sum_{i=1}^{n} \Bigg[& \exp(-c^{-1}\log(1+\mu c)) [c^{-1}(\frac{\mu}{1+\mu c}) - c^{-2}\log(1+\mu c)] \\ & \Big[[\gamma + \exp(-c^{-1}\log(1+\mu c))]^2 \Big]^{-1} I_{\{y_i=0\}} + 0 I_{\{y_i>0\}} \Bigg] \end{aligned} \tag{A.7}$$

$$\frac{\partial^2 l}{\partial \gamma^2} = \sum_{i=1}^{n} \left[(1+\gamma)^{-2} - \Big[[\gamma + \exp(-c^{-1}\log(1+\mu c))]^2 \Big]^{-1} I_{\{y_i=0\}} + 0 I_{\{y_i>0\}} \right] \tag{A.8}$$

Bibliography

Anderson, T. W. and Taylor, J. B. (1976). Strong Consistency of Least Squares Estimates in Normal Linear Regression. *The Annals of Statistics* **4**, 788–790.

Anscombe, F. J.(1949). The statistical analysis of insect counts based on the negative binomial distribution. *Biometrics* **5**, 165–173.

Barnwal, R. K. and Paul, S. R. (1988). Analysis of one-way layout of count data with negative binomial variation. *Biometrika* **75**, 215–22 .

Bohning, D., Dietz, E., Schlattmann, P., Mendonca, L., and Kirchner, U. (1999). The Zero-Inflated Poisson Model and the Decayed, Missing and Filled Teeth Index in Dental Epidemiology. *Journal of the Royal Statistical Society A* **162**, 195–209.

Bliss, C. I. and Fisher, R. A. (1953). Fitting the negative binomial distribution to biological data. *Biometrics* **9**, 176–200.

Bliss, C. I. and Owen, A. R. G. (1958). Negative binomial distribution with a common k. *Biometrika* **45**, 37–58.

Broek, J. V. D. (1995). A Score Test for Zero Inflation in a Poisson Distribution. *Biometrics* **51**, 738–743.

Breslow, N. E. (1984). Extra-Poisson variation in log-linear models. *Applied Statistics* **33**, 38–44.

Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society, Series B* **22**, 302–306.

Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* **46**, 443-459.

Cameron, A. C., and Trivedi, P. K. (2013). Regression analysis of count data. Cambridge University Press .

Casella, G. and George, E. L. (1992). Explaining the Gibbs Sampler. *The American Statistician* **46**, 167–174.

Chen, J., Hubbard, S. and Rubin, Y. (2001). Estimating the hydraulic conductivity at the south oyster site from geophysical tomographic data using Bayesian Techniques based on the normal linear regression model. *Water Resources Research* **37**, 1603–1613.

Chesher, A. (1984). Testing of neglected heterogeneity. *Econometrika* **52**, 865–872.

Cox, D. R. (1983). Some remarks on overdispersion. *Biometrika* **87**, 269–274.

Cappelli, D. P. and Mobley, C.C. (2007). Prevention in Clinical Oral Health Care. Mosby Elsevier, Philadelphia, Pa.

Dawid, A. P. and Skene, A. M. (1979). Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society. Series C* **28**, 20–28.

Dean, c. b. (1992). Testing for Overdispersion in Poisson and Binomial Regression Models. *J. Amer. Statist. Assoc.* **87**, 451–457.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society B* **39**, 1–38.

Deng, D., and Paul, S. R. (2000). Score Tests for Zero Inflation in Generalized Linear Models. *The Canadian Journal of Statistics* **87**, 451–457.

Deng, D., and Paul, S. R. (2005). Score Tests for Zero Inflation and Over Dispersion in Generalized Linear Models. *Statistica Sinica* **15**, 257–276.

Efron, B. and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika* **65**, 457–87.

Engel, J. (1984). Models for response data showing extra-Poisson variation. *Statistica Neerlandica* **38**, 159–167.

Geweke, J. (1986). Inference in the Inequality Constrained Normal Linear Regression Model. *Journal of Applied Econometrics* **1**, 117–141.

Greene, W. H. (1994). Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models. New York University , Unpublished research paper.

Gilks, W. R., and Wild, P. (1992). Adaptive Rejection Sampling for Gibbs Sampling. *Applied Statistics* **2**, 337–348.

Hall, B. H. (2000). A Note on the Bias in Herfindahl-type Measures Based on Count Data. University of California at Berkeley and NBER .

Hinde, J., and Demetrio, C. G. B. (1998). Overdispersion: Models and estimation. *Computational Statistics and Data Analysis* **27**, 151-170.

Ibrahim, J. G. (1990). Incomplete Data in Generalized Linear Model. *J. Amer. Statist. Assoc.* **85**, 765–769.

Ibrahim, J. G., and Lipsitz, S. R. (1996). Parameter Estimation From Incomplete Data in Binomial Regression When the Missing Data Mechanism Is Nonignorable. *Biometrics* **52**, 1071-1078.

Ibrahim, J. G., Lipsitz, S. R., and Chen, M.-H. (1999). Missing Covariates in Generalized Linear Models When the Missing Data Mechanism Is Nonignorable. *Journal of the Royal Statistical Society B* **61**, 173-190.

Ibrahim, J. G., Chen, M.-H., and Lipsitz, S. R. (1999). Monte Carlo EM for Missing Covariates in Parametric Regression Models. *Biometrics* **55**, 591-596.

Ibrahim, J. G., Chen, M.-H., and Lipsitz, S. R. (2001). Missing Responses in Generalized Linear Mixed Models When the Missing Data Mechanism Is Nonignorable. *Biometrika* **88**, 551-556.

Ibrahim, J. G., Chen, M.-H., and Lipsitz, S. R., and Herring, A. H. (2005). Missing Data Methods for Generalized Linear Models. *J. Amer. Statist. Assoc.* **100**, 332–346.

Jiang, X. and Paul, S. R. (2009). Analysis of covariance of zero-inflated paired count data using a zero-inflated bivariate Poisson regression model. *Calcutta Statistical Bulletin (Special Volume)* **61**, 113–124.

Kelly, B. C. (2007). Some aspects of measurement error in linear regression of astronomical data. *The Astrophysical Journal* **665**, 1489-1506.

Lawless, J. F. (1987). Negative Binomial and Mixed Poisson Regression. *The Canadian Journal of Statistics* **15**, 209–225.

Lee, A. H., Wang, K., and Yau, K. K. W. (2001). Analysis of Zero-Inflated Poisson Data Incorporating Extent of Exposure. *Biometrical Journal* **43**, 963-975.

Li, C-S , Lu, J-C , Park, J. , Kim, K., Brinkley, P. A. and Peterson, J. P. (1999). Multivariate Zero-Inflated Poisson Models and Their Applications. *Technometrics* **41**, 29–38.

Lipsitz, S. R., and Ibrahim, J. G. (1996). A conditional model for incomplete covariates in parametric regression models. *Biometrika* **83**, 916-922.

Little, R. J. A.,and Rubin, D. B. (1987, 2002, 2014). Statistical Analysis With Missing Data. New York: Wiley , 2nd ed.

Lord, D., Washington, S. P., and Ivan, J. N. (2005). Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis & Prevention* **37**, 35–46.

Lauritzen, S. L. (1995). The EM algorithm for graphical association models with missing data. *Computational Statistics & Data Analysis* **19**, 191–201.

Lipsitz, S. R., Ibrahim, J. G., and Zhao, L. P. (1999). A Weighted Estimating Equation for Missing Covariate Data with Properties Similar to Maximum Likelihood. *Journal of the American Statistical Association* **94**, 1147–1160.

Laird, N., Lange, N. and Stram, D. (1987). Maximum Likelihood Computations with Repeated Measures: Application of the EM Algorithm. *Journal of the American Statistical Association* **82**, 97–105.

Laird, N. M. and Ware, J. H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics* **38**, 963–974.

Maiti, T., and Pradhan, V. (2009). Bias Reduction and a Solution for Separation of Logistic Regression with Missing Covariates. *Biometrics* **65**, 1262–1269.

Mullahy, J. (1997). Instrumental-Variable Estimation of Count Data Models: Applications to Models of Cigarette Smoking Behavior. *The Review of Economics and Statistics* **79**, 586–593.

Minami, M. , Cody, C.E. L-. and Verdesoto, M. R-.(2007). Modeling shark bycatch: The zero-inflated negative binomial regression model with smoothing. *Fisheries Research* **84**, 210-221.

Mwalili, S. M., Lasaffre, E. and Declerck, D. (2008). The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research. *Statistical Methods in Medical Research* **17**, 123-139.

McCaughran, D. A. and Arnold, D. W. (1976). Statistical models for members of implantation sites and embryonic deaths in mice. *Toxicology and Applied Pharmacology* **38**, 325–333.

Margolin, B.H. and Kaplan, N. and Zeiger, E. (1981). Statistical analysis of the Ames salmonella/microsome test. *Proceedings of the National Academy of Sciences, U.S.A.* **76**, 3779–3783.

Manton, K. G. and Woodbury, M. A. and Stallard, E. (1981). A variance components approach to categorical data models with heterogeneous cell populations: analysis of spatial gradients in lung cancer mortality rates in North Carolina counties. *Biometrics* **37**, 259–269.

Molenberghs, G., Thijs, H. , Jansen, I., Beunckens, C., Kenward, M. G., Mallinckrodt, C., Carroll, R. J. (2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics* **5**, 445–464.

Paul, S. R. and Plackett, R. L. (1978). Inference sensitivity for Poisson mixtures. *Biometrika* **65**, 591–602 .

Paul, S. R. and Banergee, T. (1998). Analysis of Two-Way Layout of Count Data Involving Multiple Counts in Each Cell. *J. Amer. Statist. Assoc.* **93**, 1419–1429.

Paul, S. R., Jiang, X., Rai, S. N. and Balasooriya, U. (2004). Test of treatment effect in predrug and postdrug count data with zero-inflation. *Statistics in medicine* **23**, 1541–1554.

Piegorsch, W. W. (1990). Maximum likelihood estimation for the negative binomial dispersion parameter. *Biometrics* **46**, 863-867.

Prentice, R. L. (1986). Binary Regression Using an Extended Beta-Binomial Distribution, With Discussion of Correlation Induced by Covariate Measurement Errors. *J. Amer. Statist. Assoc.* **81**, 321–327.

Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997). Bayesian Model Averaging for Linear Regression Models. *J. Amer. Statist. Assoc.* **92**, 179–191.

Ridout, M, Demetrio, C. G. B. and Hinde, J. (1998). Models for count data with many zeros. *International Biometric Conference* , Cape Town.

Rubin, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys . *J. Amer. Statist. Assoc.* **72**, 538–543.

Ross, G. J. S. and Preece, D. A. (1985). The negative binomial distribution. *The Statistician* **34**, 323–336.

Sahu, S. K. and Roberts, G. O. (1999). On convergence of the EM algorithm and the Gibbs sampler. *Statistics and Computing* **9**, 55–64.

Sinha, S and Maiti, T (2007). Analysis of matched case-control data in presence of nonignorable missing exposure. *Biometrics* **64**, 106–114.

Shumway, R. H. and Stoffer, D. S. (1982). An approach to time series smoothing and forecasting using the em algorithm. *Journal of Time Series Analysis* **3**, 253–264.

Wang, K., Lee, A. H., Yau, K. K. W. and Carrivick, P. J. W. (2003). A bivariate zero-inflated Poisson regression model to analyze occupational injuries. *Accident Analysis & Prevention* **35**, 625–629.

Williamson, J. M., Lin, H-M, Lyles, R. H., and Hightower, A. W. (2007). Power Calculations for ZIP and ZINB Models. *Journal of Data Science* **5**, 519–534.

Wei, G. C. G., and Tanner, M. A. (1990). A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. of the American Statistical Association **85**, 699–704.

Xie, M., He, B., and Goh, T. N. (2001). Zero-inflated Poisson model in statistical process control. *Computational Statistics and Data Analysis* **38**, 191–201.

Zhang, C-H and Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics* **36**, 1567–1594.

Vita Auctoris

The author was born in Dhaka, Bangladesh in 1985. He obtained his B.Sc. (Honours) and M.Sc. in Applied Statistics from University of Dhaka, Bangladesh in 2008 and 2009 respectively. The author obtained his M.Sc. in Statistics from University of New Brunswick, Canada in 2011. He is currently a candidate for a Ph.D. in Statistics at the University of Windsor and will graduate in Summer 2016.