2015

Count Models With Multiple Inflations

Arvind Tripathi
University of Alabama at Birmingham

COUNT MODELS WITH MULTIPLE INFLATIONS

by

ARVIND TRIPATHI

KUI ZHANG, COMMITTEE CHAIR
INMACULADA ABAN
CHARITY MORGAN
XIAOGANG SU
HON YUEN

A DISSERTATION

Submitted to the graduate faculty of The University of Alabama at Birmingham,
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

BIRMINGHAM, ALABAMA

2015

COUNT MODELS WITH MULTIPLE INFLATIONS

ARVIND TRIPATHI

BIOSTATISTICS

ABSTRACT

The goal of this research is to develop new statistical models for the analysis of count data even if data exhibit over dispersion or under dispersion and has multiple inflated counts. Although many statistical models are available for the analysis of count data, there is no available statistical model that can address the presence of more than expected multiple counts together with over/under dispersion. In our first paper, we develop a multiple-inflation negative binomial (MINB) model and use the expectation maximization (EM) algorithm along with a numerical optimization to obtain maximum likelihood estimates. We applied the one step smoothly clipped absolute deviation (SCAD) for the variable selection. In the second paper, we develop a multiple-inflation generalized Poisson (MIGP) model and also use the EM algorithm along with a numerical optimization to obtain maximum likelihood estimates. In the third paper, we apply our novel MINB and MIGP models to data related to oral hygiene among systemic sclerosis (SSc) patients.

Based on the results from simulated data sets, we find that the MINB model, when used to analyze count data in the presence of multiple inflations and over dispersion, outperformed other existing models in terms of the average square loss (ASL). In our second paper, we obtain similar results for the MIGP model. We find that the MIGP model had the smallest ASL among the other models which can be used to model over

dispersed counts. Furthermore, in the second paper, we applied the MIGP model in real data set of social survey. In the third paper, after applying the MINB and MIGP models for analysis of oral hygiene data for systemic sclerosis (SSc) patients, we find that there is no significant association between the dental caries and SSc subtypes after adjusting for "Age" and "Income". This discovery suggests that modeling the count data without incorporating the multiple inflated counts in the analysis could provide substantially misleading results. Therefore, we strongly recommended considering the multiple inflation models when inflation in multiple counts is present.

Keywords: count model, multiple-inflation, negative binomial, generalized Poisson, dental caries, systemic sclerosis

DEDICATION

I dedicate my dissertation research to my spiritual guru, Shri Sathya Sai baba, my father, Mr. Arun Shanker Tripathi and my mother, Mrs. Amita Tripathi. Without their support and encouragement, this research would not have been possible.

ACKNOWLEDGEMENT

I feel elated in expressing my deep indebtedness to my esteemed, learned and adroit advisors Dr. Kui Zhang, and Dr. Xiaogang Su, (Associate Professor, Department of Mathematical Sciences, University of Texas at El Paso) for their proficient guidance, inspiration and keen interest throughout the progress of this research work. It was due to their kind co-operation that this research work could take its present shape. I am very thankful for their valuable suggestions, help and encouragement.

No words are available to express my profound gratefulness to Dr. David Redden, for his willingness and pains taken by him which facilitated smooth running of my academic career.

I would also like to thank my committee members: Dr. Inmaculada Aban, who collaborated well with me and gave support in all possible ways to achieve my goal, Dr. Charity Morgan, who inspired me through her research work and Dr. Hon Yuen, who not only provided me data but also provided me his prompt, extraordinary and thoughtful suggestions.

I want to extend my gratitude to Dr. George Howard for providing me financial support and encouragement to accomplish my research. I would also like to thank Dr. Leslie McClure, Dr. Hemant Tewari, Dr. Nianjun Liu, Dr. Stacey Cofield and Dr. Alfred Bartolucci whose kind help and support encouraged me to achieve my goals.

I also want to thank all the faculty members of the Department of Biostatistics for providing me insights to help me mature as a researcher and my research is the reflection of the knowledge which I absorbed from them.

Many thanks to my friends, Himel Mallick for his brotherly support and consistent encouragement and Erica Dawson for her kind support to help me in moving forward.

Last, but not the least, I want to thank my beloved wife, Saloni, who has been beside me in every step of this journey and without whom this research would not have been possible.

TABLE OF CONTENTS

Page

LIST OF TABLES

Tables *Page*

MULTIPLE- INFLATION NEGATIVE BINOMIAL MODEL WITH VARIABLE SELECTION

MULTIPLE-INFLATION GENERALIZED POISSON MODEL

NO ASSOCIATION BETWEEN DENTAL CARIES AND SYSTEMIC SCLEROSIS SUBTYPES

LIST OF FIGURES

Figure *Page*

INTRODUCTION

MULTIPLE-INFLATION GENERALIZED POISSON MODEL

NO ASSOCIATION BETWEEN DENTAL CARIES AND SYSTEMIC SCLEROSIS SUBTYPES

INTRODUCTION

Background and Motivation

The present research, namely "Count Models with Multiple Inflations," provides novel statistical models for analysis of count data with inflations at multiple counts even if data is dispersed. Many statistical models have been developed and used to describe the quantitative relationship between the response and the predictor variables. However, to select an appropriate model, the distribution of the response variable plays an important role. When the response variable is count, then we generally expect the response variable to follow a distribution specified for the counts such as a Poisson or negative binomial (NB). Often when looking at the empirical distribution of the response variable, we find some counts have a much higher frequency than expected under the Poisson or NB distribution. When only the zero count has much higher frequency than expected under the Poisson or NB distribution, then it is referred as a zero inflated (ZI) count. When two or more counts have higher frequency than expected under the Poisson or NB distribution, then we refer to them as multiple inflated (MI) counts (Su et al., 2013). Typical examples of MI counts include traffic crash data, hospitalization frequency data in health-care, DMFT/DMFS count (i.e. count of the number of the decayed, missing, filled TEETH/tooth SURFACES in a person's mouth) in dental research, and cigarette smoking data.

An example will help clarify the nature of data with the multiple inflated counts: the data archive provided by National Health and Nutrition Examination Survey

(NHANES) offers "a program of studies designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations" (mentioned in http://www.cdc.gov/nchs/nhanes.htm) (NHANES 2001-2002). Plotting any count variable (e.g. cigarettes smoked per day) from the NHANES 2001-2002* survey reveals the presence of the MI counts. More precisely, when a variable which is the response to the question "*On average, how many cigarettes {do you/does SP} now smoke per day?"* (NHANES 2001-2002) was plotted (Figure 1), the presence of the MI counts can easily be seen in the data set. Information about the inflated zeros can be obtained from number of non-smokers.

There are always some mechanisms which are responsible for the observed multiple inflated counts, and these mechanisms could be found with the proper investigation about the experiments generating the data. For example, in cigarette smoking, according to the CNN article, "Pack-a-day smokers decline" (Gardner, 2011), 23% of all smokers smoked at least one pack a day in 2007. This fact indicates that the inflated count twenty (one pack) in the cigarette smoking data set is generated by one pack a day cigarette smokers.

**The discussion about NHANES data set here is purely for demonstration purposes and no other information is used in the present research. For demonstration, the figure is created using the NHANES dataset.*

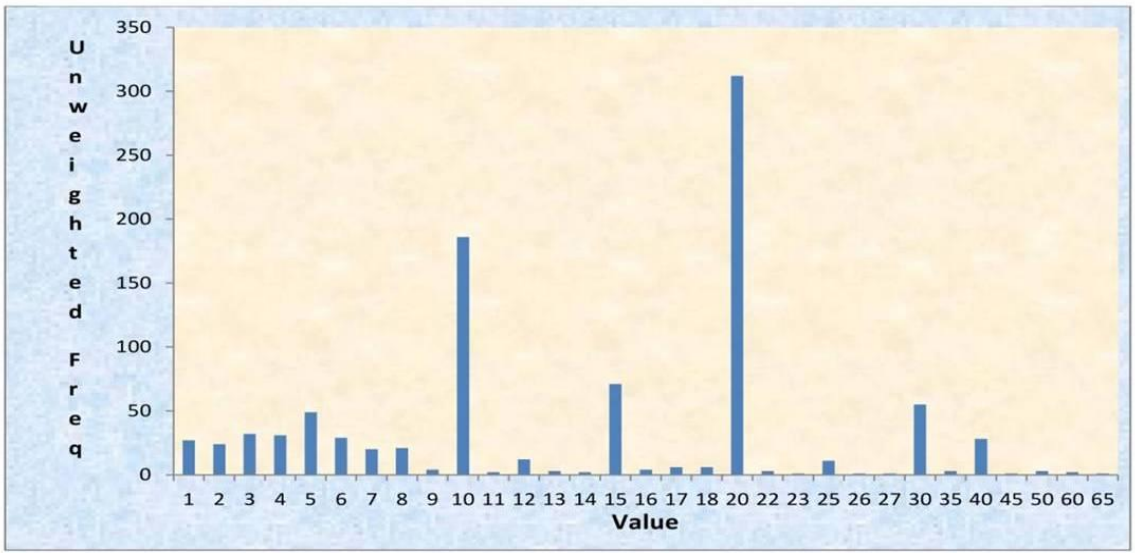


Figure 1. Histogram plot for the variable "Cigarettes smoked per day now" as given in the NHANES data set.

Similarly, the inflated count ten (half pack) in the cigarette smoking data set is generated by half pack a day cigarette smokers. Overall, 'no cigarette smokers', 'half pack a day cigarette smokers' and 'one pack a day cigarette smokers' contribute in the inflations of zeros, tens and twenties and may occur in a dataset with the certain probabilities. In the above histogram for the NHANES 2001-2002 survey data, counts of 20 (one pack of cigarettes) and 10 (half pack of cigarettes) have very high frequency relative to other counts and should be considered as inflated.

There are other types of data with multiple inflated counts. It is a widely known policy that in order for a person to receive nursing home benefits under medicare, a three-day hospital stay in the previous thirty days is required. In practice, thus it is very obvious to observe the presence of inflated zero to three counts in the thirty days hospitalization stay data, in patients with non-fatal diseases. The presence of the multiple inflated DMFT/DMFS counts has also been noticed. Preisser et al. (2012) indicated that over the past five to ten years, models for zero inflated counts have been increasingly applied to

the analysis of dental caries indices (e.g., DMFT, DMFS, etc.; dental caries result from a destructive process which not only causes decalcification of the tooth enamel but also leads to continued destruction of enamel and dentin, and cavitation of the tooth). The authors noticed that the main reason for this is linked to the broad decline in children's dental caries experience. As a result, the DMF indices more frequently generate low or even zero counts. Frequent generation of low or zero counts simply indicates the presence of the MI counts.

The statistical models that analyze count data without considering any inflated counts have long been used in many disciplines. Whenever the outcome of interest had been a count variable, investigators typically applied the Poisson regression model. However, the Poisson regression model can only be applied when the variance is approximately equal to the mean. In practice, we often sample the data from a population in which the variance of the count outcome is much different from the mean. When the variance is more than the mean, then the data is called "over dispersed". To model the over dispersed data, the NB regression model has been used much more frequently. When the variance is less than the mean then the data is called "under dispersed". The Poisson distribution is generalized in many ways, and generalized Poisson regression models are used to model the under dispersed data. The generalized Poisson models are also often found useful for over dispersed data, and they have been applied whenever data is heterogeneous.

In the above discussed models, when only the zero count is inflated, zero-inflated Poisson (ZIP), zero-inflated generalized Poisson (ZIGP) and zero-inflated negative binomial (ZINB) have been introduced and discussed below. Basically, the zero inflated

models are obtained using the concept of the mixture distribution. In fact, in a real world situation, it is not always possible to get the data from one distribution; the data may arrive from the different distributions. Lambert's (1992) work on zero-inflated Poisson (ZIP) model was motivated from such an experiment in which the counts came from the two states (perfect and erroneous) and the counts in each state followed different distributions. The perfect state followed a degenerate distribution at zero and was mixed with an untruncated Poisson distribution of the erroneous state. Eventually, the model was expressed as a mixture of two components: one component was a degenerate distribution at zero and the other component was count data following the Poisson distribution. Because no error represents the zero count, the author found more zeros in the data set were coming from the perfect state. In other words, it is observed that the perfect state was adding more zeros. However, when the number of errors in the erroneous state was modeled by a negative binomial or generalized Poisson distribution, then the zero-inflated negative binomial (ZINB) model (Greene, 1994) and the zero-inflated generalized Poisson (ZIGP) model (Famoye et al. 2006) were used, respectively.

The need for an appropriate model to deal with the presence of the multiple inflated counts had mostly been ignored till 2013, when Su et al. proposed a multiple-inflation Poisson (MIP) model to deal with the presence of multiple inflated counts. The authors extended the idea of inflation in the zero count to inflations in multiple counts. They also assumed a mixture distribution in which non-inflated counts were following a Poisson distribution and multiple inflated counts were following a degenerate distribution at the respective inflated counts. The MIP was used to model the data related to the

frequency of visits of the doctors and/or health professionals in two weeks which had two inflated counts, zero and one.

While the MIP is a valuable step, but involvement of Poisson distribution make it restrictive to equidispersed non-inflated counts and we still need an appropriate more general model to address the presence of multiple inflated counts in any situation (e.g. presence of hetrogenity). In the present research, we propose commonly used negative binomial and generalized Poisson distributions to model dispersed non-inflated counts. In particular, we propose two new models namely the multiple-inflation negative binomial (MINB) and the multiple-inflation generalized Poisson (MIGP). When there is no heterogeneity (over dispersion or under dispersion) in the data, the MINB and MIGP models will reduce to the MIP model, and when only one count zero is inflated they will reduce to the ZINB and ZIGP models respectively. Specifically, in the present research, we assume a mixture model in which discrete distributions for inflated counts is mixed with either a negative binomial or generalized Poisson distribution followed by non-inflated counts. The models presented are also more general in another sense: they incorporate the information provided by the data to model the parameters for mixing probability. In the present research, we also proposed the use of a one-step smoothly clipped absolute deviation (SCAD) method to select the important variables. The use of SCAD makes the variable selection less time consuming and more flexible in the MI models in comparison to such pre-existing methods as testing based methods, best subset selection methods and least absolute shrinkage and selection operator (LASSO). We also illustrated the application of our models (MINB and MIGP) in both simulated and real data sets. Finally, we applied our models to identify and then explore the association of

dental caries with two main subtypes of systemic sclerosis limited and diffuse cutaneous among adults. Using the above simulated and real data sets, we demonstrated that without considering multiple inflated (MI) count models, the results would have been misleading. Therefore, we strongly recommend the use of the multiple inflated count models in cases marked by the presence of significant inflation in multiple counts.

Review of Literature

Count data and its distribution have been explored in great detail for the last few centuries. The Poisson distribution and NB distribution have long been frequently used for count data. Later on, generalization of the Poisson distribution was also found useful for dispersed (over/under dispersed) counts, but the NB is preferred for the distribution of the over dispersed counts. Since the study on count data and on their distributions began centuries before, to appreciate the extraordinary contributions of the researchers' books presenting the concise history of the statistics are very useful. Dodge (2008) provided the history of the Poisson and NB distribution in his book "The Concise Encyclopedia of Statistics". The author mentioned that Pascal (1679) was the first who treated the NB distribution. Furthermore, to assess the number of times a coin should be flipped in order to get fixed number of heads Montmort (1714) used the NB distribution. Equally important, Poisson distribution carries this name due to Siméon-Denis Poisson and is the limiting case of the binomial distribution. In fact, Poisson (1837) found this distribution by considering the limits of the binomial distribution. Later on, the famous Polish statistician Ladislaus Josephovich Bortkiewicz (1898) published a book about the Poisson distribution called "The Law of Small Numbers." His book is famous for Prussian horse-kick data, he suggested that the number of soldiers killed each year by mule-kicks in the

Prussian cavalry follow a Poisson distribution. In his book, he also examined the data on child-suicides. Dodge (2008) also mentioned in his book that the use of the NB distribution as an alternative to the Poisson distribution was implemented by Student (1907), and that the application of the NB distribution is further explored by Greenwood and Yule (1920) and Eggenberger and Polya (1923).

The above details provide only a snapshot of the some important works which have been done to explore count data and their distributions in the past few centuries. Although the importance of the NB distribution to deal with real world problems could never be undermined, when the data are not over /under dispersed then the Poisson distribution comes in handy. Moreover, the need for handling over as well as under dispersed counts in a more detailed and appropriate way prompted statisticians to generalize the Poisson distribution. Consul and Jain (1973) introduced the generalized Poisson distribution with two parameters which was subsequently extensively studied by Consul (1989).

The Poisson distribution has only one parameter to estimate, i.e., mean. The variance of the Poisson distribution is same as the mean. By contrast, generalized Poisson and NB distributions have two parameters to estimate—namely mean and dispersion parameter. The variance of the NB distribution could be greater than the mean depending on the dispersion parameter. However, the variance of the generalized Poisson distribution could either be greater or less than the mean. Figure 2 illustrates the density function of the Poisson distribution for the different values of the mean parameter. Figures 3 and 4 illustrate the density function of the NB and generalized Poisson distribution for the different values of the dispersion parameter and fixed mean.

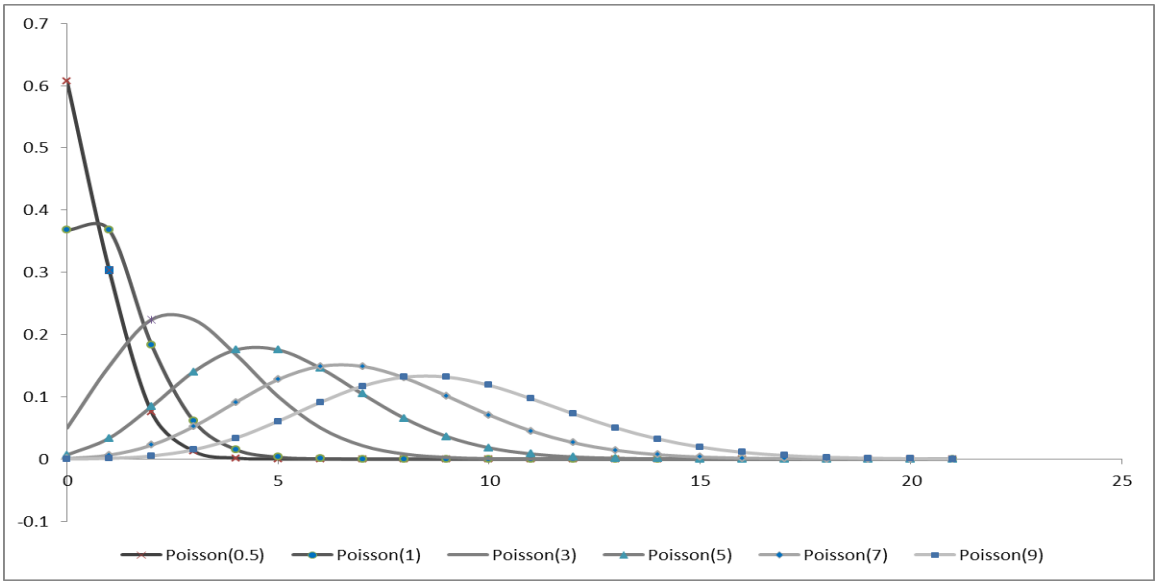


Figure 2. Probability density functions of the Poisson distributions with the means at 0.5, 1, 3, 5, 7, 9.

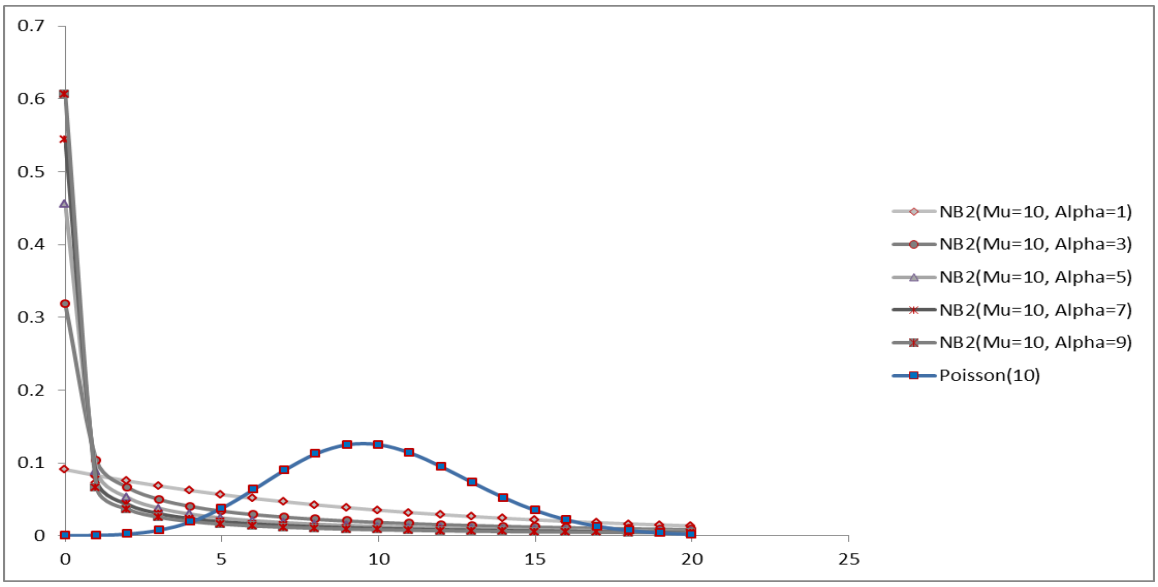


Figure 3. Probability density functions of the negative binomial distributions for a fixed value of the mean (μ), i.e. 10 and for the different values of the dispersion parameter (alpha), i.e. 1, 3, 5, 7 and 9 along with the Poisson distribution with the mean (μ), i.e. 10.

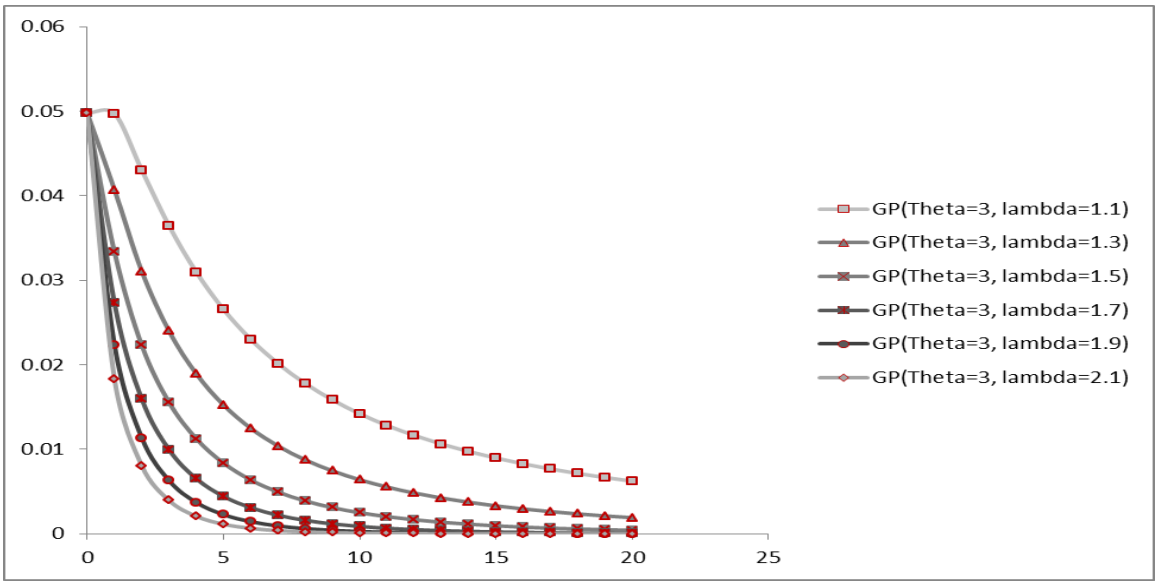


Figure 4. Probability density functions of the generalized Poisson distributions for a fixed value of the theta (mean= $\theta\nu$), i.e. 3 and for the different values of the lambda (dispersion parameter $\nu = \frac{1}{1-\lambda}$), i.e. 1.1, 1.3, 1.5, 1.7, 1.9 and 2.1.

Incorporating the above discussed distributions of the count data into the regression analysis initiated the advent of count models. For regression analysis of the count data, a unified approach developed by Nelder et al. (1972) in which dependent variables come from the different distributions of the exponential family (including count data) became very useful. Nelder et al. (1972) unified the Poisson and NB regression with other regression models and proposed generalized linear models.

Moreover, it is usually understood that in real world situations, often data come from two or more probability distributions. At first, Pearson (1894) considered the two-component normal mixture distribution. Later, the mixtures of the degenerate distribution at zero with the different count distributions became very popular and remain in common use. These are known as zero inflated (ZI) models. The idea is further extended to the multiple inflation (MI) count models using the mixture of degenerate distributions at multiple counts and Poisson distribution (Su et al., 2103). Eventually, some important MI

models are proposed in the present research which use mixture of discrete distribution and NB/ generalized Poisson distribution.

Among the above discussed models, NB2 parameterization is used for the standard NB regression model. In fact, the standard NB regression model is also termed as the NB2 model owing to the quadratic nature of its variance function (Hilbe, 2011). At the same time, NB1 parameterization, like the NB2, has a variance as linear function of the mean. NB1 model is also called linear NB model. In fact, this notion is based on the value of the exponent of the mean present in the variance function and could be generalized for any value of exponent e.g. for exponent P the parameterization is referred as NB-P (Hilbe,2011). Cameron and Trivedi (1986) were the first to explore the distinction between the NB1 and NB2 models. The parameterization of the NB2 is the traditional parameterization of the NB model and is used in the current research.

The probability density plots given in Figure 2 to 4 show that the expected number of counts under the Poisson or NB distribution may vary based on their parameters. However, the experiment produces some counts in a higher than expected frequency and some counts may not be produced at all. This leads to the concept of inflated counts and truncated counts distributions. Let us first discuss truncated distributions. For this, an example provided by Hilbe (2011) with a good explanation is used here. For the length of the stay in a hospital, as it takes value starting from 1 because a stay starts after registration and so a stay could never be zero days unless it is defined 0 for the one who never stayed in the hospital. The author further mentioned, the Poisson and NB distribution both incorporate zeros. Subsequently to exclude the zeros from the underlying distribution in accordance with the data, the distribution itself needed to be

changed. However, at the same time all the probabilities must also sum to 1, as per the definition of the density function. The zero-truncated models are proposed to accommodate the above amendments. However, the mean of the distribution is shifted towards to the left and thus results in under-dispersion. Now, as some counts may show higher frequencies than expected, the count data may also have more zeros than expected under a reference model which leads to over-dispersion. Mullahy (1986) found that there might be different mechanisms leading to these excess zeros. For example, the data may be generated from two processes with different distribution functions, where the one process generates the zero counts and the other process generates non-zero counts. Such a process leads to a model referred as a hurdle model. The author also stated that whether or not a count outcome will take a positive or zero realization is governed by a binary outcome that has a binomial probability. When the realization is positive, the hurdle is crossed and modeled with truncated-at-zero count data model. Heilbron (1994) proposed a special type of hurdle model called the zero-altered model. Zero-altered Poisson and zero-altered negative binomial models are referred as ZAP and ZANB respectively. They have also been termed overlapping models, or also zero inflated models.

Lambert (1992) found that the mechanisms leading to the excess zeros consist of two processes with different distribution functions, where one process generates the zero counts while the other process generates the counts following Poisson distribution. She proposed the zero-inflated Poisson (ZIP) regression, with an application to defects in manufacturing. Greene (1994) generalized the ZIP model to accommodate heterogeneous data and proposed zero-inflated negative binomial (ZINB) model. Hall (2000) proposed the zero-inflated binomial (ZIB) regression model and incorporated random effects into

ZIP and ZIB models; and Lee et al. (2001) provided the mixed ZIP model and accommodated the extent of individual exposure. Famoye et al. (2006) proposed a zero-inflated generalized Poisson (ZIGP) regression model and used it to model domestic violence data with too many zeros. Su et al. (2013) extended the idea of more than expected zero count by also including more than expected non-zero and/or more than one counts. They proposed multiple-inflation Poisson (MIP) model.

The present research is analogous to the MIP in that we propose a multiple-inflation negative binomial (MINB) model and a multiple-inflation generalized Poisson (MIGP) model and demonstrate their importance by using a data set related to the dental carries.

Generalized Linear Model (GLM)

Nelder and Wedderburn (1972) unified various statistical models, including linear regression, logistic regression and Poisson regression to formulate Generalized Linear Models (GLM).

The following three elements are used to generalize the linear model:

1. Any probability distribution which belongs to the exponential family.
2. A function of the predictors linear in regression parameters $\eta = X\beta$,

   where X is the design matrix and β is the vector of the regression parameters.
3. A link function g such that $g(\mu) = \eta$, where Y is a dependent variable or outcome variable and $E(Y) = \mu$.

When the outcome of interest is the count model, then the Poisson distribution becomes the obvious choice for the first element (i.e. a probability distribution from the

exponential family) and then the generalized linear model is referred as Poisson regression model. However, for the over dispersed counts most often the negative binomial (NB2) distribution is used and in such a case the generalized linear model is referred as negative binomial regression model.

Poisson regression model

The conditional probability distribution for a Poisson random variable Y_i, given the vector of covariates X_i is given by

$$Prob\left(Y_i = y_i | X_i\right) = p\left(y_i\right) = \frac{e^{-\mu_i}\left(\mu_i\right)^{y_i}}{y_i!}; y_i = 0,1,\ldots,$$

where μ_i is the parameter to be estimated, known as mean occurrence rate per unit of time, and is a function of covariates.

Therefore, we have

$$E\left(Y_i | X_i\right) = \mu_i = \exp\left(X_i^T \beta\right) \Rightarrow \log\left(E\left(Y_i | X_i\right)\right) = X_i^T \beta \ ,$$

where, β is a vector of regression parameters. The Poisson regression model is also called log-linear model because the logarithm of the conditional mean is linear in the parameters.

The Poisson model is very restrictive in the sense that it could only be applied when the conditional variance is equal to the conditional mean. Therefore, the Poisson model is not often found useful to handle real-world situations because of the presence of over or under dispersion, i.e. the variance is either more or less than the mean.

Negative Binomial Regression model

When the variance of the count data exceeds its mean then the negative binomial (NB) regression model is considered as a remedy. In the NB regression model, an

unobserved heterogeneity term for the observation *i* is introduced to allow the presence of over dispersion. Therefore, the NB regression model is more generalized than the Poisson regression model. The following parameterization of the NB model is very frequently used.

$$E(Y_i \mid X_i, \tau_i) = \mu_i \tau_i = \exp\left(X_i' \tilde{\beta}\right) \tau_i$$

where τ_i is the unobserved heterogeneity and,

$\tau_i \sim \text{gamma}(\nu, \nu) \Rightarrow E(\tau_i) = 1$ and $\text{Var}(\tau_i) = 1 \setminus \nu.$

But the dependent count variable Y_i conditional on x_i and τ_i follows Poisson distribution, i.e.,

$$P(Y_i = y_i \mid x_i, \tau_i) = \frac{\exp(-\mu_i \tau_i)(\mu_i \tau_i)^{y_i}}{y_i!}$$

and y_i follows NB distribution for given x_i

$$P(Y_i = y_i | x_i) = \frac{\Gamma(\nu + y_i)}{\Gamma(y_i + 1)\Gamma(\nu)} \left(\frac{\nu}{\mu_i + \nu}\right)^{\nu} \left(\frac{\mu_i}{\nu + y_i}\right)^{y_i}$$

where the conditional mean is given as μ_i and conditional variance is given as $\mu_i\left(1 + \frac{\mu_i}{\theta}\right)$ which is quadratic in μ_i therefore this parameterization is called NB2. The NB1 is different from the NB2 model and has a variance $\mu_i + \frac{\mu_i}{\theta}$, a linear function of μ_i.

Taking $\frac{1}{\theta} = \alpha$ variance term becomes $\mu_i(1 + \alpha\mu_i)$.

By using the fact that when we restrict a parameter in a more complex model to be zero then the result is called a nested model. Notice that when $\alpha \to 0$ the NB$\to$ Poisson model. The Poisson and the NB models are nested.

Generalized Poisson Regression Model

The Poisson model does not provide flexibility to model dispersed counts outcome due to an underlying assumption about equality in mean and variance. The generalization of the Poisson model is considered an option other than the negative binomial model. The generalization of the Poisson model can also be used to model the under dispersed (population variance is less than the population mean) counts. Moreover, NB can only be used when over dispersion (population variance is greater than the population mean) is present. Among the available generalizations, the Lagrangian-Poisson distribution has been a popular alternative (Johnson et al.,1992, p.189) to the Poisson distribution and is also known as generalized Poisson distribution (GPD)(Consul and Jain,1973). The probability density function of generalized Poisson distribution is given by:

$$P(Y=y|\theta,\lambda)=\theta(\theta+y\lambda)^{y-1}(y!)^{-1}\exp(-\theta-y\lambda),\forall\ \&\ y\notin I^{-}$$

where, $\theta>0$ and $\max\left(-1,-\frac{\theta}{m}\right)\le\lambda\le 1$, where $m\ge 4$ is the greatest positive integer satisfying $\theta+m\lambda>0$ when $\lambda<0$ (and then $P(Y=y)=0$ when $y>m$). The mean and variance of GPD are given by $\theta(1-\lambda)^{-1}$ and $\theta(1-\lambda)^{-3}$ respectively (Consul,1989), hence when sample variance is greater than sample mean, this distribution is more appropriate than Poisson distribution.

For regression purpose, we write $E(Y_i|X_i;\beta,\lambda)=\mu\left(X_i,\beta\right)=\mu_i>0$ where β a vector of regression parameters and X_i is the design matrix. Considering $\nu=\frac{1}{1-\lambda}$, the mean (i.e. $E(Y_i|X_i;\beta,\lambda)$) and the variance (i.e. $Var(Y_i|X_i;\beta,\lambda)$) of GPD can be written as μ_i and $\nu^2\mu_i$ respectively where, $\nu=\frac{1}{1-\lambda}$ represents the square root of index of dispersion, incorporating this we get the following density function

$$P\left(Y=y_i|X_i,\beta,\lambda\right)=\mu_i\left[\mu_i+\left(\nu-1\right)y_i\right]^{y_i-1}\nu^{-y_i}\left(y!\right)^{-1}\exp\left(-\frac{\left[\mu_i+\left(\nu-1\right)y_i\right]}{\nu}\right)$$

where $\mu_i>0$ *and modeled as* $\mu_i=\log\left(X\beta\right)$.

Zero- Inflated Count Models

The generalized Poisson distribution (GPD) was introduced by Consul and Jain (1973) to handle count data with over-dispersion. The generalized Poisson distribution is similar to the NB distribution in the sense that it also incorporates an extra parameter for heterogeneity or dispersion. However, the difficulty arises when the zeros in the data set are in excess. For this purpose, the zero-inflated model was introduced by Lambert (1992) in its formal sense to deal with the problem related with the excess of zeros. The zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) regression models are often used to analyze data with inflated zeros. However, hurdle models preceded zero inflated count models in the development to analyze the counts with many zeros.

Hurdle Model

The hurdle model was proposed by Mullahy (1986) using a two-stage modeling process. In the model, a binary variable that measures whether the response falls below or above the hurdle is modeled in the first stage and in the second stage a truncated model is used to model the observations above the hurdle. In the zero inflated count data, the hurdle is taken as zero. In this way, zero-inflated Poisson model could be considered a special case of the hurdle model. More precisely, a close relationship between the hurdle model and the zero inflated models can easily be observed as both are mixtures of two-components with one component a degenerate distribution at zero and the other component a count model. The second component in a hurdle model is different from the zero inflated model and follows a zero-truncated distribution whereas in a zero inflated model, it follows a non-truncated distribution.

Zero- Inflated Poisson Model

The Poisson model with mean λ and sample size n must have $ne^{-\lambda}$ expected number of zero counts (e.g. items without defects) with probability of getting the zero count (e.g. no defects) $e^{-\lambda}$.

However, in real world situations, the number of zero counts (e.g. no defects) could be more than expected. Often times, no defects are coded as zero, hence more than expected zeros means more than expected no defects. The Poisson model with more than expected zeros is called the zero-inflated Poisson (ZIP) model.

In other words, the Poisson model of count data suggests that the variance of the data should be equal to the mean. The problem arises when we find out that the data is sampled from the population having variance not equal to the mean. The two possibilities

may be there, either the variance is less than the mean, or the variance is greater than the mean. The problem that the variance is greater than the mean has been termed as over dispersion and the variance is less than the mean as under dispersion. There are many possible reasons for the over dispersion determining the specific reason requires proper investigation of the experiment that provides the data. The one possible reason which many researchers have been encountered is the excess of zeros. This problem arises when the zeros are generate by a different process in the experiment and cannot be ignored.

Lambert (1992) encountered with such a problem in which properly aligned manufacturing equipment produced zero defects, while misaligned equipments produced other defects. The zero-inflated Poisson model was first introduced by Lambert (1992) to analyze such data and then has been widely applied to health-care, economics and social sciences data that contain an excess of zeros.

In the zero-inflated Poisson (ZIP) model of Lambert (1992), the probability P is used for the possible zero observations, and $(1-P)$ for the observed Poisson (λ) random variable. The ZIP model is given by considering that the independent response variables $\mathrm{Y}=\left(\mathrm{Y}_1,\ldots,\mathrm{Y}_\mathrm{n}\right)'$ follow:

$$\mathrm{y_i} \sim \begin{cases} 0 & \textit{With probability } \mathrm{P}_i \\ \mathrm{Poisson}\left(\lambda_\mathrm{i}\right) & \textit{With probability} \left(1-\mathrm{P_i}\right) \end{cases}$$

So that,

$$\mathrm{y_i} \sim \begin{cases} 0 & \textit{With probability } \mathrm{P}_i+\left(1-\mathrm{P_i}\right)\mathrm{e}^{-\lambda_\mathrm{i}} \\ \mathrm{k} & \textit{With probability} \dfrac{\left(1-\mathrm{P_i}\right)\mathrm{e}^{-\lambda_\mathrm{i}}\left(\lambda_\mathrm{i}\right)^\mathrm{k}}{\mathrm{k!}};\mathrm{k}=1,2\ldots \end{cases}$$

Furthermore, the parameters $\lambda = (\lambda_1, \ldots, \lambda_n)'$ and $P = (P_1, \ldots, P_n)'$ are modeled as $\log(\lambda) = B\beta$ and, $\text{logit}(p) = \log\left(\frac{P}{1-P}\right) = G\tilde{\gamma}$ where B *and* G are the matrices of covariates β and γ are regression parameters.

The discussion about the number of parameters needed to be estimated provided by Lambert (1992) is worth mentioning here. As per Lambert (1992), the covariates that affect the Poisson mean of the imperfect state may or may not be the same as the covariates that affect the probability of the perfect state. When they are the same (i.e. $B = G$) and λ and P are not functionally related, then the ZIP regression requires twice as many parameters as the Poisson regression. At the other extreme, when the probability of the perfect state does not depend on the covariates and G is only a column of ones, then the ZIP regression requires only one more parameter to estimate than the Poisson regression. As per this discussion, the presence of more regression parameters to estimate in the ZI models than their non-zero-inflated counts model analogues cannot be overlooked. This underscores the importance of the variable selection in inflated count models and discussed in more detail later in the related section.

Due to the presence of two separate model components, the variable selection has become an important part of the ZIP. Not only the presence of two model components in the ZIP, but also advances in information technology have made the variable selection even more important. Modern technology has made data collection and storage easier, and thus the range of potential explanatory variables is becoming wider, including not only demographic, disease history, and medical variables but also socio-economic status, lifestyle, and genetic information. If we have k potential predictors and want each to be

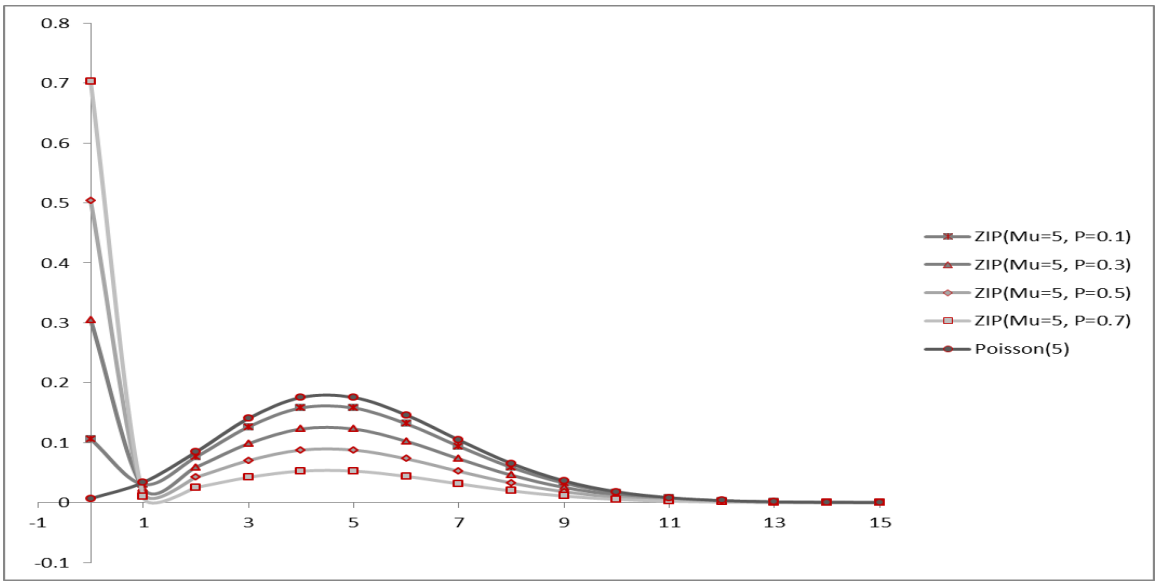
either included in or excluded from each of the two model components of the ZIP than we have 4^k possible models. However, in one model component the possible models are only 2^k. Recently for this purpose, Buu et al. (2011) proposed a new variable selection method for the ZIP model and applied it to the substance abuse field. Su et al. (2013) also used the ℓ_1 -regularization to aid in variable selection in the multiple-inflation Poisson (MIP) model. The discussion of MIP is provided below. In the present research, the method adapted by Buu et al. (2011) is implemented to aid in variable selection.

The two states assumption used by Lambert (1992) requires some further discussion. Lambert (1992) considered two states: namely, a perfect state and an erroneous state (following Poisson distribution). For estimation, she supposed that we knew which zero is coming from the perfect state and which is coming from the Poisson; that is, she introduced dummy variable Z_i such that

$$Z_i \sim \begin{cases} 1 & \text{when } Y_i \text{ is from the perfect, zero state and} \\ 0 & \text{when } Y_i \text{ is from the Poisson}(\lambda_i) \text{ state} \end{cases}$$

She used this random variable in estimating the parameters by using the EM algorithm.

Heilbron (1994) proposed the zero altered Poisson and zero altered Negative binomial regression models similar to the ZIP. But both the models (ZIP and zero altered) were developed independently. Heilbron (1994) considered a model of mixture distributions by assigning a point mass at 0 along with a positive Poisson. Figure 5 provides plots of the ZIP models and compares them for different values of mixing probabilities.



Figures 5. Zero-inflated Poisson (ZIP) models with the mean 5 and for different mixing probabilities i.e. 0.1, 0.3, 0.5 and 0.7 along with the Poisson distribution with the mean 5.

Zero-Inflated Negative Binomial (ZINB) model

The ZIP model has also been extended in several ways. When over or under dispersion exists, the one remedy might be the zero-inflated generalized Poisson (ZIGP). However, the zero-inflated negative binomial (ZINB) model is preferred when the non-inflated counts are sampled from the over dispersed population.

As mentioned earlier, the underlying assumption in the zero-inflated (ZI) models is the existence of entities in two states in which one state is called a true-zero state or the inherently safe state and another state is called a non-zero state that follows a distribution like Poisson, negative binomial (NB2) or generalized Poisson. For the ZINB model, the non-zero state follows a negative binomial (NB2) distribution i.e.:-

$$P\left(Y_i = y_i | x_i\right) = \frac{\Gamma\left(y_i + \tau^{-1}\right)}{\Gamma\left(y_i + 1\right)\Gamma\left(\tau^{-1}\right)}\left(\frac{\tau^{-1}}{\tau^{-1} + \mu_i}\right)^{\tau^{-1}}\left(\frac{\mu_i}{\tau^{-1} + \mu_i}\right)^{y_i}; y_i \geq 0$$

The ZINB model is defined as

$$y_i \sim \begin{cases} 0 & \textit{With probability } P_i \\ NB\left(\text{Mean} = \mu_i, \text{Disp.} = \tau\right) & \textit{With probability} \left(1 - P_i\right) \end{cases}$$

So that,

$$P(y_i | x_i, z_i) = \begin{cases} P_i + \left(1 - P_i\right)\left(1 + \tau\mu_i\right)^{-\tau^{-1}} & ; y_i = 0 \\ \left(1 - P_i\right) \dfrac{\Gamma\left(y_i + \tau^{-1}\right)}{\Gamma\left(y_i + 1\right)\Gamma\left(\tau^{-1}\right)} \left(\dfrac{\tau^{-1}}{\tau^{-1} + \mu_i}\right)^{\tau^{-1}} \left(\dfrac{\mu_i}{\tau^{-1} + \mu_i}\right)^{y_i} & ; y_i > 0; \end{cases}$$

where the z_i is defined as above (see the ZIP model) and the μ_i is modeled using the log-linear model and the P_i is modeled using the logistic model in the way similar to the ZIP i.e.

$$\log\left(\mu\right) = B\beta \text{ and } \text{logit}\left(p\right) = log\frac{P}{1-P} = G\gamma$$

where the B and G are the matrices of the covariates and $\tilde{\beta}$ and $\tilde{\gamma}$ are the regression parameters. The probit models instead of logit models are also used to model the P_i. The concern about the variable selection in the ZINB is the same as in the ZIP. However, to decide on the variables in each model component, the variable selection is performed before applying the zero inflated models. For example, in a study by Genuer et al. (2011) to find the relationship between Plasmodium gametocytes and their infectiousness to mosquitoes, the authors analyzed data for which the number of variables plus attendant interactions was of the order of the sample size. They first performed variable selection by applying a variable selection procedure based on the random forests score of importance and then used the ZINB on the selected variables to assess the infection prevalence.

The existence of a random effect has also been explored when there is dependency among the data due to longitudinal measures taken repeatedly on the same subjects and/or clustered design (Yau et al., 2003). Figure 6 provides plots of ZINB models and compares them for different values of mixing probabilities.

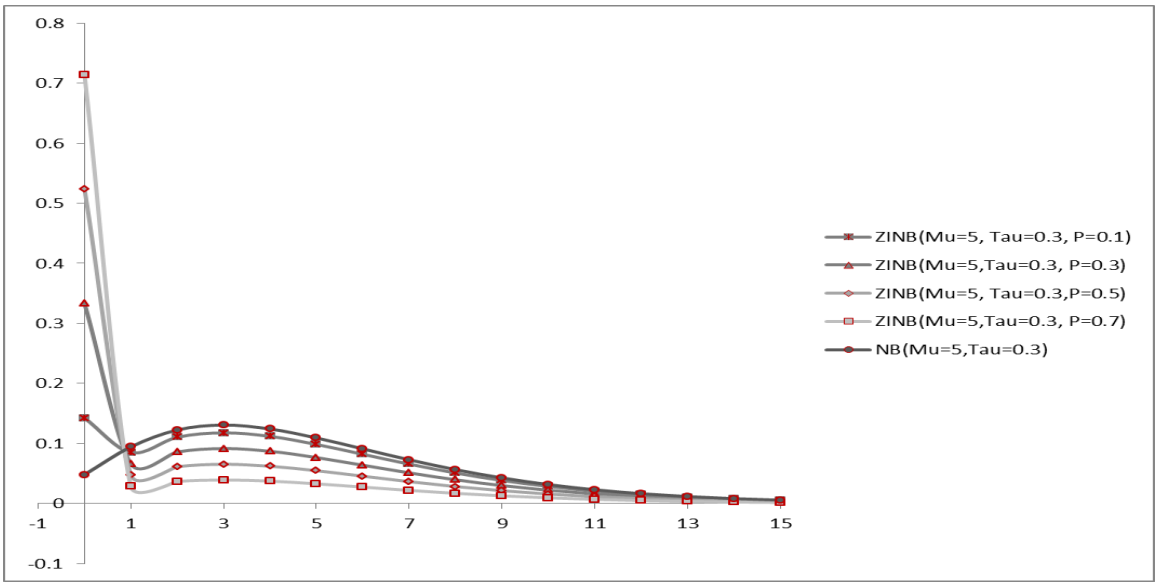


Figure 6. Zero-inflated negative binomial (ZINB) models with the mean 5, dispersion parameter 0.3 and for the different mixing probabilities, i.e. 0.1, 0.3, 0.5 and 0.7 along with the negative binomial distribution with the mean 5 and dispersion parameter 0.3.

Zero-Inflated Generalized Poisson (ZIGP) model

Famoye et al.(2006) proposed a zero-inflated generalized Poisson (ZIGP) model in a fashion similar to the ZIP and ZINB. They used following parameterization for the generalized Poisson distribution.

$$f\left(\mu_i,\alpha,y_i\right)=\left(\frac{\mu_i}{1+\alpha\mu_i}\right)^{y_i}\frac{\left(1+\alpha y_i\right)^{y_i-1}}{y_i!}\exp\left[\frac{\mu_i\left(1+\alpha y_i\right)}{\left(1+\alpha\mu_i\right)}\right];\ \ \mathrm{y_i}=0,1,2,\ldots$$

where $\mu_i=\mu_i\left(x_i\right)=\exp\left(\sum x_{ij}\beta_j\right), x_i=\left\{x_{i1}=1,x_{i2},\ldots,x_{ik}\right\}$ is the i^{th} row of the covariance matrix X and $\beta_i=\left\{\beta_1,\beta_2,\ldots,\beta_k\right\}$ is given as unknown k dimensional column vector of regression parameters, α is defined as dispersion parameter.

Similar to the other ZI models,

$$P(y_i|x_i,z_i)=\begin{cases}P_i+(1-P_i)\,f(\mu_i,\alpha;y_i); & y_i=0\\(1-P_i)\,f(\mu_i,\alpha;y_i); & y_i>0\end{cases}$$

The symbols in the above formulation are slightly changed in order to maintain the consistency across all the ZI models. The P_i and the μ_i are modeled in a fashion similar to the ZIP and ZINB, i.e. using the log linear and logistic model respectively. Figure 7 provides plots of ZIGP models and compares them for different values of mixing probabilities.

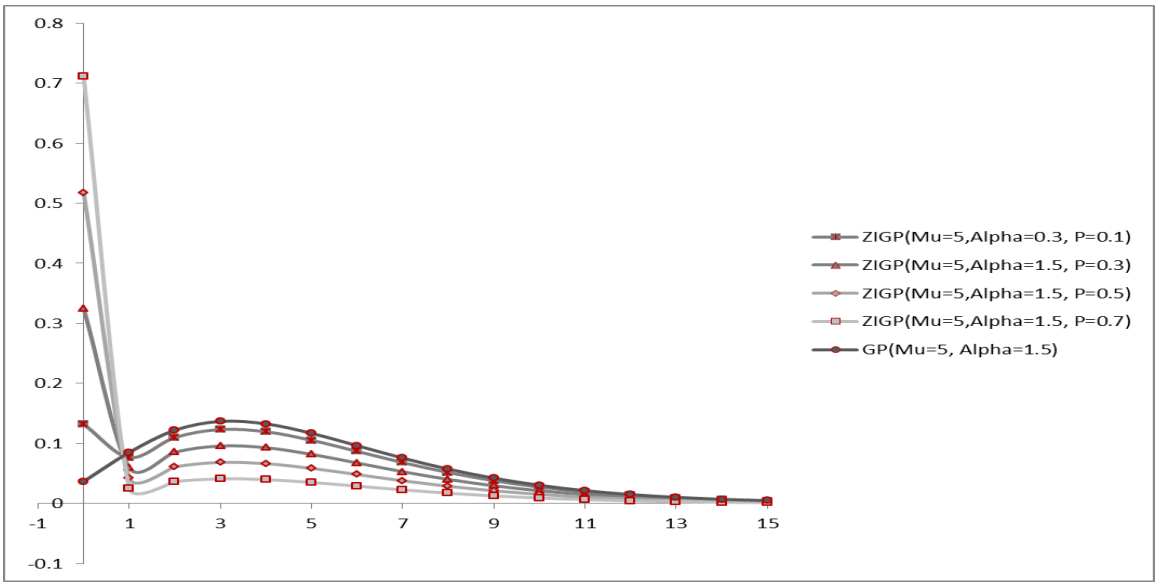


Figure 7. Zero-inflated generalized Poisson (ZIGP) models with the mean 5, dispersion parameter 0.3 and for the different mixing probabilities, i.e. 0.1, 0.3, 0.5 and 0.7 along with the plot for generalized Poisson distribution with mean 5 and dispersion parameter 0.3.

Multiple-Inflation Poisson Model (MIP)

In real world scenarios, the zero count is not the only one which is observed with the higher frequency than expected. In fact, looking at the histogram often time reveals that any count other than zero could be inflated. Even the presence of many counts as inflated counts can be easily observed. The inflation in the counts carries logic behind it

which provides information about the possible mechanism leading to these inflations. This fact was mostly overlooked until the proposal of the multiple-inflation Poisson (MIP) model. However, multiple inflation and over/under dispersion are two major problems associated with modeling the count outcome. The multiple inflations in the counts in the presence of over/under dispersion is addressed in the present research. Similar to the underlying assumption in zero inflated (ZI) models about the existence of two states i.e. perfect and erroneous, the states in the multiple inflated (MI) models are the inflated count state (i.e. a state inherently prone either to a perfect state or to a state generating some particular counts of errors) and an erroneous count state following a distribution for count data i.e. a Poisson, NB, or generalized Poisson distribution.

Su et al. (2013) proposed the MIP model. Instead of considering that count response y_i has only one value, i.e. zero as inflated, Su et al. (2013) considered that it contains a total of M inflated values. These inflated values may or may not be consecutive and denoted them as $\{0,1,\ldots,(M-1)\}$. They proposed MIP model as follows:

$$y_i \sim \begin{cases} m, & \text{with probability } p_{im} \text{ for } m = 0,1,\ldots,(M-1) \\ Poisson(\mu_i), & \text{with probability } P_{1M} \end{cases}$$

where, $\sum_{m=0}^{M} p_{im} = 1$, so that,

$$y_i \sim \begin{cases} m, \text{ with probability } p_{im} + p_{iM} e^{-\mu_i} \dfrac{\mu_i^m}{m!}; & \text{for } m = 0,1,\ldots,(M-1) \\ k, \text{ with probability } p_{iM} e^{-\mu_i} \dfrac{\mu_i^k}{k!}; & \text{for } k = 1,2,\ldots \text{ for } k \geq M \end{cases}$$

They used log-linear model to model the Poisson mean μ_i i.e.

$$\log(\mu_i) = B_i^T \beta \text{ or } \mu_i = \exp(B_i^T \beta)$$

and the cumulative logit or proportional odds model is used to model the inflated counts i.e.

$$\text{logit}\left(y_i \le m\right) = \log \frac{\Pr\left(y_i \le m\right)}{\Pr(y_i > m)} = G_i^T \gamma_1 + \gamma_{m0},$$

where the y_i represents the count response, $m = 0,1,\ldots,\left(M-1\right)$ represents the inflated values, the B_i *and* G_i associated covariate vectors and the $\beta, \gamma_1 and \gamma_{m0}$ the vectors for regression parameters of loglinear model, slope parameters and intercept parameters of cumulative logit model respectively.

They also expressed the MIP model by using $\left(M+1\right)$ states. It should be noticed that Lambert (1992) used two such states which were referred as perfect state and erroneous state. For M inflated counts, i.e. $0,1,\ldots,\left(M-1\right)$, the authors took state 0 for $y_i = 0$, state 1 for $y_i = 1$, and so on, state $\left(M-1\right)$ for $y_i = \left(M-1\right)$ and at state M for y_i following Poisson(μ_i).

They introduced dummy variables z_{im} such that $z_{im} = 1$ if y_i is from the m^{th} state and 0 otherwise, for $\text{m} = 0,1,\ldots,\text{M}$ and $\text{i} = 1,\ldots,\text{n}$. Thus, $\sum_{\text{m}} \text{z}_{\text{im}} = 1$ and $\text{z}_{\text{im}}\text{z}_{\text{im'}} = 0$ for any $\text{m} \neq \text{m'}$ .Conditioning on these dummy variables, they provided the distribution of y_i as

$$\text{y}_\text{i} \mid \text{z}_{\text{io}},\ldots,\text{z}_{\text{iM}} \sim \begin{cases} \Pr\left(\text{y}_\text{i} = 0 \mid \text{z}_{\text{i0}} = 1\right) = 1, \\ \ldots, \\ \Pr\left(\text{y}_\text{i} = \text{M}-1 \mid \text{z}_{\text{i(M-1)}} = 1\right) = 1, \\ \text{y}_\text{i} \mid \left(\text{z}_{\text{iM}} = 1\right) \sim \text{Poisson}\left(\mu_\text{i}\right) \end{cases}$$

They also proposed a different model formulation obtained via a mixture of a discrete distribution over all inflated values $\{0,1,\ldots,(M-1)\}$ and a Poisson distribution, where the mixture probability is supplied by a Bernoulli model.

$$y_i \sim \begin{cases} \text{Discrete}\{(0,\ldots,(M-1)); (p'_{i0},\ldots,p'_{i(M-1)}\} & \textit{with prob.}\, 1-p_{iM} \\ \textit{Poisson}(\mu_i) & \text{with prob.}\, p_{iM} \end{cases}$$

where, $\sum_{m=0}^{M-1} p'_{im} = 1$ and other parameters are modeled as follows:

$$\begin{cases} \log(\mu_i) = B_i^T\beta \text{ or } \mu_i = \exp(B_i^T\beta), \\ \text{logit}(\Pr\{y_i \le m\}) = \log\dfrac{\Pr\{y_i \le m\}}{\Pr\{M-1 \ge y_i > m\}} = G_i^T\gamma_1 + \gamma_{m0} \text{ for} m = 0,\ldots(M-2), \\ \text{logit}(p_{iM}) = H_i^T\gamma_2 + \gamma_{M0} \end{cases}$$

The zero-inflated models are called the ZIP models when the non-inflated counts follow a Poisson distribution having variance equal to the mean. However, equidispersion (i.e. population variance is equal to mean) is found very restrictive and the NB distribution and generalized Poisson distribution are used frequently for over/under dispersed data to model the non-zero state and the models are referred as the ZINB and ZIGP models. Similarly, multiple-inflation negative binomial (MINB) and multiple-inflation generalized Poisson (MIGP) models are proposed in this research to model the data when non inflated counts are present without equidispersion and with over/under dispersion. It is recommended in the present research that multiple inflated models should be applied if inflation is present in the data to avoid misleading results.

It worth mentioning here, the Poisson, NB and generalized Poisson distributions were not always the only choice as one of the mixing distributions in zero inflated models. The possibility of mixing other distributions has also been explored. Mixing zeros with

right censored continuous distributions to model survival data (Farewell, 1986 and Meeker, 1987) and mixing zeros with a gamma distribution to model rainfall data (Feuerverge,1979) are a few such examples. Hence, multiple-inflated models have also potential to be explored for the choices other than Poisson, NB and generalized Poisson distributions as mixing distributions.

As mentioned earlier in the above sections related with the ZIP and ZINB, the variable selection again plays a very important role in the application of such models. Su et al. (2012) used ℓ_1 regularization to select the important variables. They minimized the following:

$$\min_{\theta} -\mathrm{L}(\theta) + \sum_{j} \lambda_j \left|\theta_{1j}\right|$$

where $L(.)$ is a log-likelihood function and $\lambda_j \geq 0$ are the tuning parameters, which after local quadratic approximation of log-likelihood changes into the quadratic programming problem.

Variables Selection

As mentioned earlier, the zero inflated (ZI) models have two separate model components. Lambert (1992) used a log-linear model and a logistic model as two separate model components. The predictors for these model components may be the same and could also be entirely different. Due to the two model components, there are at the most $4^{\#\text{ of predictors}}\left(2^{2\#\text{ of predictors}}\right)$ possible models (instead of at the most $2^{\#\text{ of predictors}}$ possible models in non-zero-inflated analogs of ZI models). Therefore, it becomes even more difficult in ZI models than in their non-zero-inflated analogues to find the important variables associated with the outcome variable. Consequently, variable selection has

always been an issue whenever the zero inflated model is used. Selecting the important variables which are associated with the outcome variable has been the topic of interest from a very long time. The recent developments in this area have made variable selection very straight forward and less time consuming.

Testing based Methods

Some widely used variable selection methods are based on the sequential hypothesis testing. These methods are in much use because they provide a simple way to choose the predictors which are significantly associated with the outcome variable. But these methods somehow lack the well justified theoretical background.

Stepwise selection method. The forward, backward and bidirectional sequential testing procedures are among the most widely used variable selection procedures and are in very common use. Efroymson (1960) proposed this as a widely used algorithm. In the forward selection method, we start with the null model and then sequentially add predictors if they are significant on the basis of t or F tests based on pre-specified Type-I error rate. Conversely, in the backward selection we start with the full model and then remove the non-significant predictors one by one taking the pre specified Type-I error rate for the significance level. The bidirectional is the combination of the backward and the forward selection methods. In this, at each step we test which variable should be included or excluded at the pre specified probabilities for selection and staying in the model. But when the variables are large, this method is computationally prohibitive.

Best Subset Selection. The best subset selection method based on all subset comparisons is an alternative to the stepwise selection method. This method is based on an algorithm known as branch-and-bound or leaps-and-bound (LB) algorithm provided

by Furnival and Wilson (1974). Furnival and Wilson solved the following problem i.e. for all integer k such that $1 \le k \le p$

$$\min_{x} \left\| y - \beta x \right\|_2^2$$

subject to: $\left\| \beta \right\|_0 = k$

where $\left\| . \right\|_2^2$ is a l_2 norm, i.e., it denotes the sum of squares of the elements of a vector.

$\left\| . \right\|_0 = \text{card}()$ (i.e., cardinality) is a l_0 quasi-norm i.e. the number of non-zero entries in a vector, y is a response variable, β is a regression coefficient, x is a predictor variable and p is the number of predictors. Unfortunately, the best subset selection method is not very useful because it gets infeasible when the number of predictors gets large.

Penalized likelihood Criteria

The algorithm provided by Furnival and Wilson (1974) is connected with many widely used model selection methods. These methods basically address the following optimization problem:

$$\min_{x} \left\| y - \beta x \right\|_2^2 + \lambda_0 \left\| \beta \right\|_0$$

where $\left\| . \right\|_2^2$ is a ℓ_2 norm i.e. denotes the sum of squares of the elements of a vector. $\left\| . \right\|_0$ is a ℓ_0 quasi norm, i.e. the number of non-zero entries in a vector and the λ_0 an algorithmic parameter.

Since in the linear models setting, minimizing the residual sum of squares is equivalent to maximizing the log-likelihood, the algorithm given above can also be written in the following way:

$$\text{minimize } L\left(X, Y, \hat{\beta}\right) - c\left(\hat{\beta}\right)$$

where $L(.)$ is a log-likelihood function and $\hat{\beta}$ is the estimate of the regression coefficient.

Akaike Information Criterion (AIC). Akaike (1974), a Japanese statistician, first provided the Akaike Information Criterion and then Sugiura (1978) proposed it for the linear regression models. For the AIC, we optimize the below objective function:

$$\left\|y-\beta x\right\|_2^2+\lambda_0\left\|\beta\right\|_0$$

by taking λ_0 as $2\sigma^2$, where σ^2 is an unbiased estimate of the common variance of the random error (Ni et al., 2006).

The more frequently appearing form of the AIC is

$$AIC=-2L\left(X,Y,\hat{\beta}\right)+2p$$

where p is the number of parameters in the model.

In order to select the best model, we calculate AIC values for each model with the same data set, and the "best" model is the one which has minimum AIC value. The problem associated with this method is that the value of AIC depends on y (data), this leads to uncertainty in the model selection.

Bayesian information criterion (BIC) or Schwarz criterion (SBC). The Bayesian information criterion (BIC) was first proposed by Gideon E. Schwarz (1978). For BIC, we optimize the following algorithm:

$$\left\|y-\beta x\right\|_2^2+\lambda_0\left\|\beta\right\|_0$$

by taking λ_0 as $2\sigma^2\log n$, where σ^2 is an unbiased estimate of the common variance of the random error (Ni et al., 2006).

The more frequently appearing form of the BIC is:

$$\mathrm{BIC} = -2\mathrm{L}\left(\mathrm{X}, \mathrm{Y}, \hat{\beta}\right) + \mathrm{pln}\left(\mathrm{n}\right)$$

where p is the number of parameters in the model and n is the total number of observations.

In order to select the best model, we calculate BIC values for each model with the same data set, and the "best" model is the one which has minimum BIC value. The difference between AIC and BIC is that AIC penalizes the number of parameters less strongly than BIC. The situation where the number of observations is much lower than the number of exposure variables arises in many fields, including genetics data analysis. In this situation, traditional methods (regression) do not work. Therefore, the researchers started using "penalized regression" to come up with a better solution. Penalized regression uses a penalty term and depending upon the penalty term many different methods are proposed including LASSO and ridge regression. Typically, LASSO is found superior to ridge regression, but if there are groups of variables among which the pairwise correlations are very high, then LASSO tends to select only one variable from the group and does not care which one is selected. LASSO is provided by Tibshirani (1996).

Before LASSO, it would be worth mentioning about ridge regression. Ridge regression was proposed by Hoerl and Kennard (1970) who suggested the use of the following algorithm:

$$\text{minimize} \left\| y - \beta x \right\|_2^2 + \lambda \sum_{j=1}^{p} {\beta_j}^2$$

where λ is a tuning parameter.

Ridge regression provides continuous shrinkage and achieves better prediction performance than ordinary least square (OLS) through a bias-variance trade-off (biased estimates with lower variance). Ridge regression keeps all the predictors in the model, and therefore cannot produce a parsimonious model.

In LASSO, we use the following algorithm

$$\text{minimize}\left\|y-\beta x\right\|_2^2+\lambda\left\|\beta\right\|_1$$

where $\left|\left|\beta\right|\right|$ is a ℓ_1 norm i.e. the sum of the absolute values of the elements of the vector and λ is a tuning parameter.

LASSO uses ℓ_1 penalty instead of ℓ_0 penalty. Tibshirani (1996) has demonstrated that LASSO is more stable and accurate than traditional variable selection methods such as best subset selection. A remarkable property of LASSO is that it can automatically achieve variable selection by shrinking some coefficients to zero. Due to the shrinkage, only important variables remain in the model.

Bridge Regression. Frank and Friedman (1993) proposed bridge regression, a broad class of the penalized regression method which can be obtained by minimizing

$$\text{minimize}\left\|y-\beta x\right\|_2^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|^{\alpha}$$

$$\text{such that } \lambda,\alpha\geq 0$$

where the α is known as a concavity parameter that controls the concavity of the function, and λ is known as a tuning parameter. Bridge regression includes subset selection with $\alpha = 0$, LASSO with $\alpha = 1$ and ridge regression with $\alpha = 2$ as special cases. Bridge regression does both the variable selection (when $0 \leq \alpha \leq 1$) and shrinks the

coefficients (when α> 1). Frank and Friedman (1993) did not solve the bridge regression for any given α> 0, but they pointed out that optimizing the parameter α is desirable.

Smoothly Clipped Absolute Deviation (SCAD). It is soon realized that variable selection with a convex penalty is achieved at the cost of bias in the estimators. The need of the penalty function to get unbiasedness, continuity and sparsity in the estimators is always being realized. Fan and Li (2001) introduced the SCAD penalty which not only aids in the variable selection, but also provides the estimates with the above three (i.e. unbiasedness, continuity and sparsity) properties. The SCAD penalty is non convex in nature and is given as follows:

$$p_{\lambda}\left(|\beta|\right)=\begin{cases}\lambda|\beta| & if\ \ 0\le|\beta|<\lambda\\ -\dfrac{|\beta|^2-2a\lambda|\beta|+\lambda^2}{2(a-1)} & if\ \lambda\le|\beta|<a\lambda\\ \dfrac{(a+1)\lambda^2}{2} & if\ |\beta|\ge a\lambda\end{cases}$$

The conventional convex optimization algorithm is found not very suitable to optimize the objective function with such a singular and non-convex penalty. Hence, Fan and Li (2001) suggested local quadratic approximation (LQA) of the penalty function. However, Fan and Li (2001) realized that LQA suffers the drawbacks similar to the backward variable selection and the variable once removed could never be considered again. Later, Zou and Li (2008) proposed the local linear approximation (LLA) of the likelihood function with non-concave penalty and found this algorithm as computationally efficient as LASSO and suggested the use of efficient algorithm like least angle regression (LARS) to solve it.

Along with a SCAD variable selection other methods such as minimax concave penalty (MCP) proposed by Zhang (2010) has been studied extensively and is also considered to satisfy the three properties, namely unbiasedness, continuity and sparsity.

Recently, the new variable selection method, namely subtle uprooting has been proposed by Su (2014). The subtle uprooting method approximates the cardinality involved in the information criterion with a smooth function and provides variable selection in one step. The author used modified BFGS algorithm for non-convex optimization in variable selection with subtle uprooting.

In the previous discussion, the catalog of methods and models with some brief commentary on their historical development is provided to set forth the key terms. The following paragraphs offer a brief explanation of the issues which motivate the whole work and contents of the subsequent chapters showing how they are organized to address these issues.

As per the discussion provided above, the none of the models except MIP address the issue of multiple inflated counts, however inclusion of Poisson distribution make it restrictive to equidispersed non-inflated counts and need of an appropriate more general model to address the presence of multiple inflation in heterogeneous counts remained unanswered.

In the discussion about MIP model, Su et al. (2013) provided many examples with the multiple inflated counts such as traffic data where, for example, the number of monthly car crashes on high speed roadway segments is mostly zeros, ones, and twos; the number of insurances that are of different types and of different policies; the number of hospitalization days in healthcare applications, and the cigarette smoking data from the

"National Health and Nutrition Examination Survey (NHANES)" in which number of cigarettes smoked per day is dominated by zeros, tens (half pack of cigarettes) and twenties (one pack of cigarettes).

In all aforementioned examples (Su et al., 2013), there is also a strong possibility of getting over dispersed non-inflated counts. For example, we found that in most of cigarette smoking data sets from NHANES, the non-inflated counts are over dispersed and should not be modeled with the Poisson distribution. There are no existing models that address such a situation (i.e., the presence of the multiple inflations in over dispersed counts) precisely and appropriately. We propose a multiple-inflation negative binomial (MINB) model in first paper and multiple-inflation generalized Poisson (MIGP) model in second paper to deal with such situations. When there is no heterogeneity (over dispersion or under dispersion) in the data, the MINB and MIGP models will reduce to the MIP model, and when only one count zero is inflated they will reduce to the ZINB and ZIGP models respectively. In this way MIGP model proposed in the second paper provides more general framework and can be applied to model any count data even if it is heterogeneous and have multiple inflated counts. In the first paper, we also applied the one-step smoothly clipped absolute deviation (SCAD) method to select the important variables. The use of SCAD makes the variable selection less time consuming and more flexible in the MI models (with the three model components) in comparison to such pre-existing methods as testing based methods, best subset selection methods and LASSO. We also illustrated the application of our models (MINB and MIGP) in simulated data sets. Finally in third paper, we applied our models to identify and then explore the association of dental caries with two main subtypes of systemic sclerosis limited and

diffuse cutaneous among adults. In the third paper, using the data related with dental caries among systemic sclerosis patients, we demonstrated that without considering proposed novel multiple inflated (MI) count models the results would have been misleading. Therefore, we strongly recommend the use of the multiple inflated count model in cases marked by the presence of significant inflation in multiple counts.

MULTIPLE-INFLATION NEGATIVE BINOMIAL MODEL WITH VARIABLE SELECTION

by

ARVIND TRIPATHI, KUI ZHANG AND XIAOGANG SU

In preparation for Statistica Sinica

Format adapted for dissertation

1. ABSTRACT

In modeling count data, difficulties often arise when the outcome variable is not only dispersed but also has more than one inflated count. Analogous to the multiple-inflation Poisson (MIP) model (Su et al., 2013), we propose a multiple- inflation negative binomial (MINB) regression model by using a mixture of a cumulative logit model and negative binomial model, whereas the mixing probabilities are formulated with a logistic regression. An EM algorithm is developed to obtain maximum-likelihood estimates. Moreover, the smoothly clipped absolute deviation (SCAD) with some important modifications is adapted to aid in the variable selection issues with the MINB model. The simulated data are used to assess the performance of the proposed model and compare it with the other available competitive count models.

2. INTRODUCTION

The Poisson and negative binomial (NB) are two commonly used models for the analysis of count data. However, the count data may have certain counts in higher frequencies than was expected under a Poisson or NB distribution so cannot be readily modeled by either Poisson or NB model. We refer to such counts as inflated counts or inflation in counts. There might be certain reasons for getting inflated counts, and understanding these reasons requires a proper investigation of the data. But, even if we find a legitimate reason behind the inflation in the certain counts, accurate analysis remains a challenge. Although, if only zero count is inflated, zero inflated (ZI) models including the zero-inflated Poisson (ZIP) model (Lambert, 1992) and the zero-inflated negative binomial (ZINB) model (Greene, 1994) can be used for the analysis. The ZI models are based on the concept of mixture distributions. Lambert's (1992) work on the ZIP model was motivated from an experiment in which the counts were from two states and the counts in each state followed different distributions: the counts from one state followed a degenerate distribution at zero and was mixed with an untruncated Poisson distribution of the counts from the other state. However, when the counts from the other state were dispersed and followed the NB distribution, then ZINB model can be used to obtain more accurate results.

The possibilities of getting other than zero or more than zero counts as inflated exist in many real world data sets. Su et al. (2013) found that the inflation in zeros and ones in data related to a healthcare study on the frequency of medical visits, and they proposed the multiple-inflation Poisson (MIP) model as a mixture of the Poisson distribution and a discrete distribution. They provided many other examples with the

multiple inflated counts such as traffic data where, for example, the number of monthly car crashes on high speed roadway segments is mostly zeros, ones, and twos. For this example, the authors also mentioned about the transportation literature that have concerned about the two states assumption of the ZIP model along with the multiple state crash process as recommended solution and suggested the use of the multiple-inflation count model, i.e., MIP model as a solution. The authors also listed several other examples with the multiple inflated counts including the number of insurances that are of different types and of different policies; the number of hospitalization days in healthcare applications; and the cigarette smoking data from the "National Health and Nutrition Examination Survey (NHANES)" in which number of cigarettes smoked per day is dominated by zeros, tens (half pack of cigarettes) and twenties (one pack of cigarettes).

In all aforementioned examples (Su et al., 2013), there is also a strong possibility of getting over dispersed non-inflated counts. For example, we found that in most of the cigarette smoking data sets from NHANES, the non-inflated counts are over dispersed and should not be modeled with the Poisson distribution. There is no existing model that addresses such a situation (i.e., the presence of the multiple inflations in over dispersed counts) precisely and appropriately. We propose a multiple-inflation negative binomial (MINB) model to deal with such situations.

In the MINB model, we use a mixture of the discrete distribution and NB distribution to model the multiple inflated counts and the non-inflated counts, respectively. We further use the logit model to model the mixing probability. Hence, the proposed MINB model consists of three model parts to model the multiple inflated counts, non-inflated counts and the mixing probability. To select appropriate covariates in each

of the three models, we use one-step smoothly clipped absolute deviation (SCAD) which is used and recommended by Buu et al. (2011) for zero inflated count data for variable selection due to its high specificity, sensitivity, exact fit and lowest estimation error in comparison to least absolute shrinkage and selection operator (LASSO).

This paper is organized as follows. In Section 3, we propose a multiple-inflation negative binomial (MINB) model including its mixture model representation, identifiability, and (over and under) dispersion. In Section 4, we derive the maximum likelihood (ML) estimation along with EM algorithm. In Section 5, we describe one-step smoothly clipped absolute deviation (SCAD) method for the variable selection in the MINB model. Sections 6 to 8 consist of evaluation and comparison of the MINB model with the other models in simulated data.

3. MULTIPLE-INFLATION NEGATIVE BINOMIAL (MINB) MODEL

For the MINB model, we considered a data set with n independent observations, namely $\{(y_i, \mathbf{X}_i) : i = 1, \cdots, n\}$ where, y_i represents the count outcome and $\mathbf{X}_i$ the associated predictor vector for observation i. Suppose that the count outcome has M inflated values. These inflated counts have a natural order in it and thus can be arranged either in an ascending or descending order. Without loss of generality, we arrange the inflated counts in the ascending order and denote them as c_i, where i represents the order of the count. Hence, the set of the inflated counts is denoted as $\mathcal{I} = \{c_1, \ldots, c_M : c_1 \le c_2 \le \cdots \le c_M\}$. For example, in the case of the one inflated count, e.g., zero (i.e. zero-inflated count data), we have $M = 1$ and $\mathcal{I} = \{c_1 \; i.e. \; 0\}$, in the case of the two inflated counts, e.g., 10 and 20, we have $M = 2$ and $\mathcal{I} = \{c_1, c_2\}$ $i.e. \{10, 20\}$ and with

the three inflated counts, say 10, 15 and 20, we have $M = 3$ and $\mathcal{I} = \{c_1, c_2, c_3\}$ *i.e.* $\{10, 15, 20\}$.

3.1 Model Specification

The multiple-inflation negative binomial (MINB) model is given by:

$$y_i \sim \begin{cases} \text{Discrete}\{c_1, c_2, \ldots, c_M ; p_{i1}, p_{i2}, \ldots, p_{iM}\} & \text{with probability } (1-\phi_i) \\ \text{Poisson}(\mu_i, \tau) & \text{with probability } \phi_i \end{cases} \quad ..(0.1)$$

where, $\tau \sim \text{Gamma}(\nu, \nu)$ with $E(\tau) = 1$ and $\text{var}(\tau_i) = 1/\nu$. After incorporating this, the equation (1.1) is further written as:

$$y_i \sim \begin{cases} \text{Discrete}\{c_1, c_2, \ldots, c_M ; p_{i1}, p_{i2}, \ldots, p_{iM}\} & \text{with probability } (1-\phi_i) \\ \text{NB}(\mu_i, \nu) & \text{with probability } \phi_i \end{cases}$$

Essentially, the MINB assumes a mixture model of a discrete distribution over the inflated values in $\mathcal{I}$ and a negative binomial i.e. $\text{NB}(\mu_i, \nu)$ distribution, where the mixing probability is guided by ϕ_i. While other forms of defining the NB components, i.e., $\text{NB}(\mu_i, \nu)$ are available, the Poisson-Gamma model form has been conveniently used in the MINB.

Other involved parameters $\{p_{im}, \phi_i, \mu_i : m = 1, \ldots, M \text{ and } i = 1, \ldots, n\}$ are further specified with three regression models. To proceed further, we first introduce a $M+1$-dimensional dummy variable vectors $\boldsymbol{\delta}_i = (\delta_{im})$ and $\delta_{im} \in R^{(M+1)}$ with $m = 1, 2, \ldots, M+1$ for each unit-i such that:

$$\delta_{im} = \begin{cases} 1\{y_i = c_m\} & \text{for } m = 1, 2, \ldots, M \\ 1\{y_i \notin I\} & \text{for } m = M+1 \end{cases}.$$

Since the inflated values in $\mathcal{I}$ are naturally ordered, it is convenient to apply a cumulative logit or proportional odds regression model as follows:

$$\log \frac{\Pr\left(y_i \le c_m | \delta_{i(M+1)} = 0\right)}{\Pr\left(c_m < y_i \le c_M | \delta_{i(M+1)} = 0\right)} = \gamma_{0m} + \boldsymbol{g}_i^T \boldsymbol{\gamma}_1 \ \forall m = 1, 2, \ldots, M-1.$$

Secondly, a logistic regression is used to model the mixing probability

$$\log\left(\frac{\phi_i}{1-\phi_i}\right) = \boldsymbol{h}_i^T \boldsymbol{\alpha}.$$

Finally, a NB regression model is used to model the regular counts in $\mathrm{NB}(\mu_i, \nu)$ with

$$\log\left(\mu_i\right) = \boldsymbol{b}_i^T \boldsymbol{\beta}$$

In the above specifications, $\{\mathbf{g}_i, \mathbf{h}_i, \mathbf{b}_i\}$ are the covariate vectors of appropriate dimensions; they consist of selected components from $\mathbf{X}_i$. For convenience, we denote $\boldsymbol{\gamma} = \left(\gamma_{01}, \gamma_{02}, \ldots, \gamma_{0(M-1)}, \boldsymbol{\gamma}_1^T\right)^T$ and let $\boldsymbol{\theta} = \left(\boldsymbol{\alpha}^T, \boldsymbol{\gamma}^T, \boldsymbol{\beta}^T, \boldsymbol{\nu}\right)^T$ be the vector that collects all the parameters involved in the model.

From the model specified in the above equations, it follows that $\mu_i = \exp\left(\boldsymbol{b}_i^T \boldsymbol{\beta}\right)$ and $\phi_i = expit\left(\boldsymbol{h}_i^T \boldsymbol{\alpha}\right)$ where the expit function is given by $\mathrm{expit}(t) = \{1 + \exp(-t)\}^{-1}$. As per the cumulative model, we have

$$p_{im} = \Pr\left(y_i = c_m \mid \delta_{i(M+1)} = 0\right)$$

$$= \begin{cases} expit\left(\gamma_{01} + \boldsymbol{g}_i^T \boldsymbol{\gamma}_1\right) & for\ m = 1 \\ expit\left(\gamma_{0m} + \boldsymbol{g}_i^T \boldsymbol{\gamma}_1\right) - expit\left(\gamma_{0(m-1)} + \boldsymbol{g}_i^T \boldsymbol{\gamma}_1\right) & for\ m = 2, 3, \ldots, M-1 \\ 1 - expit\left(\gamma_{0(M-1)} + \boldsymbol{g}_i^T \boldsymbol{\gamma}_1\right) & for\ m = M \end{cases} \quad \text{.....(0.2)}$$

when $M = 2$, we have $p_{i1} = \Pr\left\{y_i = c_1 | \delta_{i(M+1)} = 0\right\}$ and $p_{i2} = \Pr\left\{y_i = c_2 | \delta_{i(M+1)} = 0\right\}$ only.

3.2 Dispersion

The presence of the multiple inflated counts also induces either the over dispersion or under dispersion (Su et al., 2013). The condition for the over and under dispersion along with the expected value and variance for the MINB model is provided as follows:

3.2.1 Expectation

According to the MINB model, the mean of y_i is given by:

$$E(y_i) = E(E(y_i|\delta_i) = (1-\phi_i)\sum_{m=1}^{M} c_m p_{im} + \phi_i \exp(\boldsymbol{b}_i^T \boldsymbol{\beta})$$

This result can be used for prediction purpose.

3.2.2 Variance

The variance of y_i is given by:

$$Var(y_i) = Var(E(y_i \mid \delta_i)) + E(Var(y_i \mid \delta_i))$$

Or we can write it as,

$$Var(y_i) = E(y_i) + (1-\phi_i)\left[\sum_{m=1}^{M} c_m^2 p_{im} - (1-\phi_i)\left(\sum_{m=1}^{M} c_m p_{im}\right)^2 - \sum_{m=1}^{M} c_m p_{im}\right]$$
$$+\phi_i \exp(\boldsymbol{b}_i^T \boldsymbol{\beta})\left(\frac{\exp(\boldsymbol{b}_i^T \boldsymbol{\beta})}{\nu} + \exp(\boldsymbol{b}_i^T \boldsymbol{\beta}) - \phi_i \exp(\boldsymbol{b}_i^T \boldsymbol{\beta})\right)$$

$$= E(y_i) + (1-\phi_i)\{\sum_{m=1}^{M} c_m(c_m-1)p_{im} - (1-\phi_i)\sum_{m=1}^{M} c_m^2 p_{im}^2\} + \phi_i \exp(2\mathbf{b}_i^T \boldsymbol{\beta})(\frac{1}{\nu} + (1-\phi_i))$$

$$Var(y_i) = E(y_i) + (1-\phi_i)\left[\sum_{m=1}^{M} c_m^2 p_{im} - (1-\phi_i)\left(\sum_{m=1}^{M} c_m p_{im}\right)^2 - \sum_{m=1}^{M} c_m p_{im}\right]$$
$$+\phi_i \exp(2\boldsymbol{b}_i^T \boldsymbol{\beta})\left(\frac{1}{\nu} + (1-\phi_i)\right)$$

The over-dispersion can be obtained if

$$(1-\phi_i)\left[\sum_{m=1}^{M} c_m^2 p_{im} - (1-\phi_i)\left(\sum_{m=1}^{M} c_m p_{im}\right)^2 - \sum_{m=1}^{M} c_m p_{im}\right] + \phi_i \exp\left(2\boldsymbol{b}_i^T \boldsymbol{\beta}\right)\left(\frac{1}{\nu} + (1-\phi_i)\right) > 0$$

Therefore, the condition for the over-dispersion is

$$\phi_i \exp\left(2\boldsymbol{b}_i^T \boldsymbol{\beta}\right)\left(\frac{1}{\nu} + (1-\phi_i)\right) > (1-\phi_i)\left[(1-\phi_i)\left(\sum_{m=1}^{M} c_m p_{im}\right)^2 + \sum_{m=1}^{M} c_m p_{im} - \sum_{m=1}^{M} c_m^2 p_{im}\right]$$

Similarly, the condition for the under dispersion is

$$\phi_i \exp\left(2\boldsymbol{b}_i^T \boldsymbol{\beta}\right)\left(\frac{1}{\nu} + (1-\phi_i)\right) < (1-\phi_i)\left[(1-\phi_i)\left(\sum_{m=1}^{M} c_m p_{im}\right)^2 + \sum_{m=1}^{M} c_m p_{im} - \sum_{m=1}^{M} c_m^2 p_{im}\right]$$

The point should be noticed that the left side of the inequality depends upon the parameters of the NB model and the mixing probability, while the right part of the inequality depends upon the parameters of the discrete distribution and the mixing probability. For the NB model, i.e., on taking $c_m = 0$ and $\phi_i = 1$, the above expression provides the over dispersion for the $\nu > 0$ and the under dispersion for the $\nu < 0$.

3.4 Identifiability

The identifiability is a very important property of a model in order to make an inference. Moreover, to correctly identify the model, it must provide different distributions for different sets of parameters. If the two sets of the different parameters are giving the same distribution of the observations, then the model is called not identifiable. We could also impose certain restrictions in order to achieve identifiability for a model; the set of these restrictions or requirements is called the identification conditions. A model can be identifiable, partially identifiable or non-identifiable (unidentifiable). If a model is non-identifiable, but it is possible to find the true values of a certain subset of the model parameters then the model is called partially identifiable.

For mixture models, if any one of the models in the class (family) can be uniquely characterized then the mixture of models is considered identifiable.

Teicher (1961) has defined the identifiability of the mixture models as follows. Assuming a measurable subset $\mathcal{R}_1^{\kappa'}$ of the Euclidean κ' -space $\mathcal{R}^{\kappa'}$ and $\mathcal{F} = \{\mathbf{F}(\mathbf{x};\boldsymbol{\alpha}), \alpha \in \mathcal{R}_1^{\kappa'}\}$ where $\mathbf{F}(\mathbf{x};\boldsymbol{\alpha})$ (measurable on the product space of the $\mathbf{x}$ and $\boldsymbol{\alpha}$) is a cumulative distribution function in the variable $\mathbf{x}$ for each $\alpha \in \mathcal{R}_1^{\kappa'}$ then for any non-degenerate κ'-dimensional c.d.f. $\mathbf{G}$ such that the induced Lebesgue-Stieltjes measure of $\mathbf{G}$ (i.e. μ_G) assigns measure one to $\mathcal{R}_1^{\kappa'}$, the c.d.f. $\mathcal{H}(\mathbf{x}) = \int_{\mathcal{R}_1^{\kappa'}} F(\mathbf{x};\boldsymbol{\alpha})\, \mathbf{dG}(\boldsymbol{\alpha})$ is called a mixture or more precisely a $\mathbf{G}$ mixture of $\mathcal{F}$. The family $F(\mathbf{x};\boldsymbol{\alpha})$, $\boldsymbol{\alpha} \in \mathcal{R}_1^{\kappa'}$ and $\mathbf{G}$ are known as the Kernel of the mixture and a mixing distribution function respectively. The mixture $\mathcal{H}$ is said to be identifiable if there is unique $\mathbf{G}$ yielding $\mathcal{H}$. According to Tallis et al. (1982), after imposing the restriction on $\mathcal{R}_1^{\kappa'}$ that it consists of a finite number of elements, i.e., $\mathcal{R}_1^{\boldsymbol{\kappa}'} = \{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots \boldsymbol{\alpha}_n\}$, the $\mathcal{H}(\mathbf{x})$ can be written as $\mathcal{H}(\mathbf{x}) = \sum_{i=1}^{n} \omega_i \mathbf{F}_i(\mathbf{x})$ where $\sum_{i=1}^{\eta} \omega_i = 1$. They suggested that the above mixture is identifiable if and only if $F = \{\mathbf{F}_1(\mathbf{x}), \mathbf{F}_2(\mathbf{x}), \ldots, \mathbf{F}_\mathbf{n}(\mathbf{x})\}$ is linearly independent.

Proposition 3.1

The MINB model is identifiable (See supplementary material).

4. MAXIMUM LIKELIHOOD ESTIMATION

Analogous to the MIP (Su et al., 2013), the MINB model is fully specified with the three regression models, as mentioned above. To write down its likelihood function, we note that

$$\pi_{ic} = \Pr\{y_i = c\} = \begin{cases} (1-\phi_i)\, p_{im} + \phi_i . \Psi_i & for\ c \in I\ and\ m = 1,2,\ldots,M \\ \phi_i . \Psi_i & for\ c \notin I \end{cases}$$

where ψ_i denotes the probability density function associated with the $\mathrm{NB}(\mu_i, \nu)$, namely,

$$\psi_i = \frac{\Gamma(y_i + \nu)}{y_i ! \Gamma(\nu)} \left[\frac{\nu}{\nu + \mu_i}\right]^{\nu} . \left[\frac{\mu_i}{\nu + \mu_i}\right]^{y_i}$$

The likelihood function $l(\theta)$ of MINB model is then given by:

$$\ell(\boldsymbol{\Theta}) = \prod_{i=1}^{n} \left[\prod_{m=1}^{M} \left\{ (1-\phi_i)\, p_{im} + \phi_i \Psi_i \right\}^{\delta_{im}} \right] . \left(\phi_i \Psi_i\right)^{\delta_{i(M+1)}}$$

The corresponding log-likelihood $L(\theta)$ is

$$L(\Theta) = \log \ell(\Theta) = \sum_{i=1}^{n} L_i(\Theta)$$

$$= \sum_{i=1}^{n} \left\{ \left[\sum_{m=1}^{M} \delta_{im} \log\left\{ (1-\phi_i)\, p_{im} + \phi_i \Psi_i \right\} \right] + \delta_{i(M+1)} \log\left(\phi_i \Psi_i\right) \right\}$$

Su et al. (2013) suggested the use of quasi-Newton method to find the estimates of the MIP model because of the complicated form of the hessian matrix. They used BFGS quasi-Newton method and suggested the use of L-BFGS when the number of parameters is very large. However, we used the default method provided in R optim() function which is Nelder-Mead simplex algorithm. We found that it is a commonly used technique for nonlinear optimization when the derivatives are not known, even though it is a heuristic approach and could also some time converge in non-stationary points.

4.1 EM Algorithm

The expectation-maximization (EM) algorithm is an iterative method for finding the maximum likelihood estimates of the parameters of the mixture models. Though the EM algorithm was used previously by some authors, it is more formally introduced by Dempster et al. in 1977. The EM algorithm consists of the iteration of the two steps, namely E-Step and M-step, and breaks the likelihood into the components which can be easily optimized but is slow in convergence. In each iteration, the E-step involves taking the conditional expectation over the complete-data log likelihood conditional upon the observed data and the parameters that we use to evaluate the expectation. The complete data consists of unknown data or latent variable Z and the observed data X and the current parameter estimates. The M-step in each iteration only requires maximizing the expectation of the log likelihood function. For the EM algorithm, we introduce a random variable $\mathbf{Z}$ consist of Z_{im} and $Z_{i(M+1)}$ as follows, i.e., for given i

$$Z_{im} = \begin{cases} 1 & \text{if } \mathrm{y_i} = \mathrm{c_{im}} \forall \mathrm{c_{im}} \in I : m = 1, 2, \ldots, \mathrm{M} \\ 0 & \text{if } \mathrm{y_i} \in \mathrm{NB} \end{cases}$$

and

$$Z_{i(M+1)} = \begin{cases} 1 & \text{if } \mathrm{y}_i \in \mathrm{NB} \\ 0 & \text{otherwise} \end{cases} ; \text{ notice that } Z_{i(M+1)} = 1 - \sum_{i=1}^{m} Z_{im}$$

For given m, the random variables Z_{im} and $Z_{i(M+1)}$ are partially observable; Z_{im}'s value is known only if $y_i \neq c_{im}$ and it is unknown otherwise. The subscripts i and m for the Z are used in the above description to incorporate the i^{th} random variable corresponding to the m^{th} inflated count.

When the cumulative logistic and log-linear models have no parameters in common and given an estimate of Z_{im}, an iteration of the EM algorithm reduces to the estimation of the $\boldsymbol{\alpha}$ using the logistic regression model and taking the $Z_{i(M+1)}$ as the response variable and the estimation of the $\boldsymbol{\beta}$ and ν by using the NB regression model with y_i as the response variable and the $Z_{i(M+1)}$ as a weight.

For given m, the complete-data likelihood associate with the $Y_c = (Y_i; Z_{im})$ and the parameter space Θ is obtained as follows:

$$f(Y_c \mid \boldsymbol{\Theta}) = f(Y_i, Z_{im} \mid \boldsymbol{\Theta}),$$

where $\boldsymbol{\Theta} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\nu}\}$

$$f\left(Y_i, Z_{im} \mid \boldsymbol{\Theta}\right) = f(Y_i \mid Z_{im}, \boldsymbol{\Theta}) f(Z_{im} \mid \boldsymbol{\Theta})$$

Hence, the complete data likelihood can be written as follows:

$$\ell = \prod_{i=1}^{n}\prod_{m=1}^{M} f\left(Y_i, Z_{im} \mid \boldsymbol{\Theta}\right) = \prod_{i=1}^{n}\prod_{m=1}^{M} \left\{ f\left(Y_i, Z_{im} = 1 \mid \boldsymbol{\Theta}\right)\right\}^{Z_{im}} \left\{ f\left(Y_i, Z_{im} = 0 \mid \boldsymbol{\Theta}\right)\right\}^{Z_{i(M+1)}}$$

or

$$= \prod_{i=1}^{n}\prod_{m=1}^{M} \left\{ f\left(Y_i \mid Z_{im} = 1, \boldsymbol{\Theta}\right) f\left(Z_{im} = 1 \mid \boldsymbol{\Theta}\right)\right\}^{Z_{im}} \left\{ f\left(Y_i \mid Z_{im} = 0, \boldsymbol{\Theta}\right) f\left(Z_{im} = 0 \mid \boldsymbol{\Theta}\right)\right\}^{Z_{i(M+1)}}$$

$$\Rightarrow \ell = \prod_{i=1}^{n}\prod_{m=1}^{M} \left\{ P_{im}\left(1-\phi_i\right)\right\}^{Z_{im}} \left\{ \Psi_i \phi_i \right\}^{Z_{i(M+1)}}$$

Therefore, the log-likelihood of complete data is given by:

$$L_c\left(\Theta \mid Y_c\right) = \sum_{i=1}^{n}\left[\sum_{m=1}^{M} \left\{ Z_{im}\left(\log p_{im}\right) + Z_{im} \log\left(1-\phi_i\right)\right\} + Z_{i(M+1)}\left(\log \phi_i + \log \Psi_i\right)\right]$$

Noticing the fact that,

$$\sum_{m=1}^{M} Z_{im} \log\left(1-\phi_i\right) + Z_{i(M+1)} \log\left(\phi_i\right) = Z_{i(M+1)} \boldsymbol{h}_i^T \boldsymbol{\alpha} - \log\left(1+\exp\left(\boldsymbol{h}_i^T \boldsymbol{\alpha}\right)\right)$$

$$L_c\left(\boldsymbol{\Theta} \mid Y_c\right) = \sum_{i=1}^{n} [\{\sum_{m=1}^{M} Z_{im} \left(\log p_{im}\right)\} - \log\left(1+\exp\left(\boldsymbol{h}_i^T \boldsymbol{\alpha}\right)\right) + Z_{i(M+1)} \left(\boldsymbol{h}_i^T \boldsymbol{\alpha} + \log \Psi_i\right)]$$

where,

$$\psi_i = \frac{\Gamma\left(y_i+\nu\right)}{y_i!\Gamma\left(\nu\right)} \left[\frac{\nu}{\nu+\mu_i}\right]^{\nu} \cdot \left[\frac{\mu_i}{\nu+\mu_i}\right]^{y_i} \text{ and } \log\left(\mu_i\right) = \boldsymbol{b}_i^T \boldsymbol{\beta}$$

This reduces into

$$= \sum_{i=1}^{n} [\{\sum_{m=1}^{M} Z_{im} \left(\log p_{im}\right)\} + Z_{i(M+1)} \left(\log \Psi_i\right) + Z_{i(M+1)} \boldsymbol{h}_i^T \boldsymbol{\alpha} - \log\left(1+\exp\left(\boldsymbol{h}_i^T \boldsymbol{\alpha}\right)\right)]$$

Noticing that $\ell(\boldsymbol{\Theta} \mid Y_c)$ is linear in Z_{im} for m=1,$\cdots$, M and $Z_{i(M+1)}$, therefore at $(k+1)^{th}$ iteration of the algorithm, the E- step consists of replacing Z_{im} for m=1,$\cdots$, M and $Z_{i(M+1)}$ by their conditional expectations, given observed data y_i and parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ *and* ν.

More detailed expression of the EM algorithm at the $k+1^{th}$ iteration involves the following steps:

4.1.1 E-Step

Computation of the conditional expectation of the Z_{im} for $m = 1, ..., M$ and $Z_{i(M+1)}$ denoted by $\hat{Z}_{im}^{(k)}$ and $\hat{Z}_{i(M+1)}^{(k)}$ (i.e. $E[Z_{im} \mid Y_i, \boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\gamma}^{(k)}, \nu^{(k)}] = \hat{Z}_{im}^{(k)}$ and $E[Z_{i(M+1)} \mid Y_i, \boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\gamma}^{(k)}, \nu^{(k)}] = \hat{Z}_{i(M+1)}^{(k)}$) respectively.

$$\begin{cases} \hat{Z}_{im}^{(k)} = \dfrac{\left(1-\left(\phi_i \mid \hat{\boldsymbol{\alpha}}^{(k)}\right)\right) p_{im}\delta_{im}}{\left(1-\left(\phi_i \mid \hat{\boldsymbol{\alpha}}^{(k)}\right)\right) p_{im} + \left(\phi_i \mid \hat{\boldsymbol{\alpha}}^{(k)}\right)(\psi_i \mid \hat{\boldsymbol{\beta}}^{(k)}, \hat{\nu}^{(k)})} & \text{if } y_i \in I \text{ and } m = 1,2\ldots,M \\ \hat{Z}_{i(M+1)}^{(k)} = 1 - \sum_{m=1}^{M} \hat{Z}_{im}^{(k)} \end{cases}$$

So that,

$$Q(\boldsymbol{\Theta} \mid \hat{\boldsymbol{\Theta}}^{(k)}) = E(l_c(\boldsymbol{\Theta} \mid Y_c) \mid Y, \hat{\boldsymbol{\Theta}}^{(k)}) = \sum_{i=1}^{n} Q_{1i}(\boldsymbol{\gamma} \mid \hat{\boldsymbol{\Theta}}^{(k)}) + \sum_{i=1}^{n} Q_{2i}(\boldsymbol{\beta}, \nu \mid \hat{\boldsymbol{\Theta}}^{(k)}) + \sum_{i=1}^{n} Q_{3i}(\boldsymbol{\alpha} \mid \hat{\boldsymbol{\Theta}}^{(k)})$$

where,

$$Q_{1i}\left(\gamma \mid \hat{\boldsymbol{\Theta}}^{(k)}\right) = \sum_{m=1}^{M} \hat{Z}_{im}^{(k)} \left(\log p_{im}\right);$$

$$Q_{2i}(\boldsymbol{\beta}, \nu \mid \hat{\boldsymbol{\Theta}}^{(k)}) = \hat{Z}_{i(M+1)}^{(k)} log\{\frac{\Gamma(y_i + \nu)}{y_i!\Gamma(\nu)}\left(\frac{\nu}{\nu + exp(\mathbf{b}_i^T \boldsymbol{\beta})}\right)^{\nu}\left(\frac{\exp(\mathbf{b}_i^T \boldsymbol{\beta})}{\nu + exp(\mathbf{b}_i^T \boldsymbol{\beta})}\right)^{y_i}\}$$

and

$$Q_{3i}(\boldsymbol{\alpha} \mid \hat{\boldsymbol{\Theta}}^{(k)}) = \hat{Z}_{i(M+1)}^{(k)} \mathbf{h}_i^T \boldsymbol{\alpha} - log(1 + \exp \mathbf{h}_i^T \boldsymbol{\alpha})$$

4.1.2. M-Step

In the *M*-step, we estimate $\boldsymbol{\beta}$ and ν, by maximizing $Q_{2i}(\boldsymbol{\beta}, \nu \mid \hat{\boldsymbol{\Theta}}^{(k)})$ which is the log-likelihood for a weighted NB regression of y_i on $\boldsymbol{b}_i$ with the weights $\hat{Z}_{i(M+1)}^{(k)}$. In the *M*-step, we also estimate $\boldsymbol{\gamma}$ and in this step, we maximize $Q_{1i}(\boldsymbol{\gamma} \mid \hat{\boldsymbol{\Theta}}^{(k)})$ which has a form similar to the log-likelihood obtained from a proportional odds model with responses $\hat{Z}_{im}^{(k)}$. The conjugate gradients method (with commonly used Polak-Ribiere formula) does not store a matrix and is used for the fast and straight forward maximization. However to get the estimates of $\boldsymbol{\alpha}$, in the *M*-step, we maximize $Q_{3i}(\boldsymbol{\alpha} \mid \hat{\boldsymbol{\Theta}}^{(k)})$, which has a form

similar to the log-likelihood obtained from an un-weighted binomial logistic regression of $\hat{Z}_{i(M+1)}^{(k)}$ on $\mathbf{h_i}$.

The iterations are repeated until $\|\hat{\Theta}^{(k+1)} - \hat{\Theta}^{(k)}\|$ is sufficiently small, $K = 0, 1,$ in order to supply the initial values for the parameters, we used the similar strategy which was used by Su et al. (2013), i.e., a NB model to the data with no inflated counts and a cumulative logit model to the inflated counts is fitted. After fitting these models, the resulting estimates are used as the initial values for the $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ respectively. The resulting estimates of a logistic regression model (using a dichotomous variable with 0 for inflated and 1 for non-inflated counts) are used as the initial value for the $\boldsymbol{\alpha}$. However, these values are not directly supplied as the initial values to the EM algorithm and used as the initial values for the numerical optimization first. The estimates obtained after the numerical optimization are supplied as initial values to the EM algorithm. Hence, we adopted a three step procedure to obtain maximum likelihood estimates. Precisely, this procedure to fit the MINB model consists of the following three steps i) getting initial values by fitting the NB model to the non-inflated counts, cumulative logit model to the inflated counts and logit model to dichotomous variables discussed above; ii) using these values as the initial values for running the default method for optim(), i.e. derivative-free optimization routine, Nelder-Mead simplex algorithm to provide the initial values for the EM algorithm; iii) using the EM algorithm to update the estimates. Do et al. (2008) has mentioned, in most of the non-concave optimization methods, the EM algorithm provides assurance only for convergence to a local optimum of the objective function. The derivative-free optimization routine Nelder-Mead simplex algorithm which is default in

optim() is used to get the initial values for the EM algorithm to facilitate its convergence to the global optimum.

The Hessian matrix obtained from R function optim() is used to approximately estimate the variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ (*i.e.* $\boldsymbol{\Sigma} = Cov\left(\hat{\boldsymbol{\theta}}\right)$) via observed Fisher's information matrix as $\hat{\boldsymbol{\Sigma}} = \{-\boldsymbol{H}\}^{-1}$. However, the optim() function uses the finite difference method to approximate the hessian matrix.

5. VARIABLE SELECTION VIA ONE-STEP SCAD METHOD

The presence of more than one model components in the zero and multiple-inflation count models makes the variable selection an important issue and thus addressed in the present research.

Due to the three model components, there are at the most $8^{\#\text{ of predictors}}$ ($2^{3^{\#\text{ of predictors}}}$) possible models (instead of at the most $2^{\#\text{ of predictors}}$ possible models in non-inflated counts analogs of MI models); therefore, it becomes even more difficult in MI models to find the important variables associated with the outcome variable.

The forward, backward and bidirectional sequential testing procedure is among the most widely used variable selection procedures and are very common in use. However, when the variables are large, this method is computationally prohibitive. The best subset selection method based on all subset comparisons is an alternative to stepwise selection method. However, this method is also not very useful because it gets infeasible when the number of predictors gets large. Tibshirani (1996) has demonstrated that LASSO is more stable and accurate than traditional variable selection methods such as best subset selection. A remarkable property of the LASSO is that it can automatically

achieve variable selection by shrinking some coefficients to zero. However it is soon realized that the variable selection with convex penalty (e.g. LASSO) is achieved at the cost of biasedness in the estimators. The need of the penalty function to get unbiasedness, continuity and sparsity in the estimators has long been realized. Fan and Li (2001) introduced the SCAD penalty which not only aids in variable selection but also provides the estimates with the above three (i.e. unbiasedness, continuity and sparsity) properties.

The conventional convex optimization algorithm is found not very suitable to optimize the objective function with a singular and non-convex penalty like SCAD. Hence, Fan and Li (2001) suggested local quadratic approximation (LQA) of the penalty function. However, Fan and Li (2001) realized that LQA suffers the drawbacks similar to the backward variable selection and the variable once removed could never be considered again. Later, Zou and Li (2008) proposed the local linear approximation (LLA) of the likelihood function with non-concave penalty and found this algorithm as computationally efficient as LASSO, and they suggested the use of an efficient algorithm like least-angle regression (LARS) to solve it. Buu et al. (2011) used one step SCAD for the variable selection in the ZIP model after local linear approximation (LLA) of the likelihood function. To aid variable selection, we are also using one step SCAD for the variable selection in the multiple inflated counts setting.

After adding the penalty term in the likelihood, we maximize the following penalized likelihood function for the MINB model:

$$Q(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}) = \ell(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}) - n\sum_{j=1}^{p_1} p_{\zeta_j}(\boldsymbol{\alpha}_j) - n\sum_{k=1}^{p_2} p_{\eta_k}(\boldsymbol{\beta}_k) - n\sum_{l=1}^{p3} p_{\kappa_l}(\boldsymbol{\gamma}_l),$$

where p_{ζ_j} , p_{η_k} and p_{κ_l} are penalty functions with tuning parameters ζ_j, η_k *and* κ_l respectively. We used different tuning parameters for different regression parameters and did not apply any penalty on the intercepts of the model components. However, the above formulation can be used for several choice of penalty functions but in the present research we are using SCAD penalty proposed by Fan and Li(2001).The SCAD penalty is given by

$$p_\tau(\boldsymbol{\theta}) = f(x) = \begin{cases} \tau|\boldsymbol{\theta}| & ,0 \le |\boldsymbol{\theta}| < \tau \\ -\left(\dfrac{\left(|\boldsymbol{\theta}|^2 - 2c\tau|\boldsymbol{\theta}| + \tau^2\right)}{2(c-1)}\right) & ,\tau < |\boldsymbol{\theta}| \le c\tau \\ \dfrac{(c+1)\tau^2}{2} & ,|\boldsymbol{\theta}| > c\tau \end{cases}$$

The c =3.7 is suggested by Fan and Li (2001) and used here. One-step sparse estimation method was proposed by Zou and Li (2008) and was used for the maximization of penalized likelihood of zero inflated count models by Buu et al. (2011).

The algorithm provided by Zou and Li (2008) for the implementation of one step sparse estimation for SCAD penalty (Buu et al. 2011) is used in the current research. For this, the likelihood function is locally approximated at the initial value $\left(\boldsymbol{\alpha}^{(0)}, \boldsymbol{\beta}^{(0)}, \boldsymbol{\gamma}^{(0)}\right)$ and after taking initial values as maximum likelihood estimates, i.e., $\left(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right)$, the likelihood function is reduced into the following form:

$$\ell\left(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right) + \frac{1}{2}[\left(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}\right)', \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)', \left(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}\right)'] \Delta^2 \ell\left(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right) \begin{bmatrix} \left(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}\right)' \\ \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)' \\ \left(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}\right)' \end{bmatrix}$$

This form is obtained because at the maximum likelihood estimates the first order derivative of the likelihood function is zero $\left(i.e.\ \Delta\ell\left(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}\right)=0\right)$. Also, the local linear approximation (LLA) to the penalty function is given as below:

$$p_{\boldsymbol{\zeta}_j}(\boldsymbol{\alpha}) \approx p_{\boldsymbol{\zeta}_j}\left(\left|\hat{\boldsymbol{\alpha}}_j\right|\right) + p'_{\boldsymbol{\zeta}_j}\left(\left|\hat{\boldsymbol{\alpha}}\right|\right)\left(|\boldsymbol{\alpha}_j| - \left|\hat{\boldsymbol{\alpha}}_j\right|\right),\ for\ \boldsymbol{\alpha}_j \approx \hat{\boldsymbol{\alpha}}_j$$

$$p_{\boldsymbol{\eta}_k}(\boldsymbol{\beta}) \approx p_{\boldsymbol{\eta}_k}\left(\left|\hat{\boldsymbol{\beta}}_k\right|\right) + p'_{\boldsymbol{\eta}_k}\left(\left|\hat{\boldsymbol{\beta}}\right|\right)\left(|\boldsymbol{\beta}_k| - \left|\hat{\boldsymbol{\beta}}_k\right|\right),\ for\ \boldsymbol{\beta}_k \approx \hat{\boldsymbol{\beta}}_k\ and$$

$$p_{\boldsymbol{\kappa}_l}(\boldsymbol{\gamma}) \approx p_{\boldsymbol{\kappa}_l}\left(\left|\hat{\boldsymbol{\gamma}}_k\right|\right) + p'_{\boldsymbol{\kappa}_l}\left(\left|\hat{\boldsymbol{\gamma}}\right|\right)\left(|\boldsymbol{\gamma}_l| - \left|\hat{\boldsymbol{\gamma}}_l\right|\right),\ for\ \boldsymbol{\gamma}_l \approx \hat{\boldsymbol{\gamma}}_l$$

The first derivative of SCAD penalty is given by:

$$p'_{\tau}(\boldsymbol{\theta}) = f(x) = \begin{cases} \tau\ sign(\boldsymbol{\theta}) & ,0 \le |\boldsymbol{\theta}| < \tau \\ \left(\dfrac{sign(\boldsymbol{\theta})(c\tau - |\boldsymbol{\theta}|)}{(c-1)}\right) & ,\tau < |\boldsymbol{\theta}| \le c\tau \\ 0 & ,|\boldsymbol{\theta}| > c\tau \end{cases}$$

using locally approximated penalty function, the penalized likelihood reduces in the following form:

$$Q(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}) \approx \ell\left(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}\right) + \frac{1}{2}\left[\left(\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}}\right)', \left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)', \left(\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}}\right)'\right][\Delta^2\ell\left(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}\right)]\begin{bmatrix}\left(\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}}\right)' \\ \left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)' \\ \left(\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}}\right)'\end{bmatrix}$$

$$-n\sum_{j=1}^{p_1} p'_{\zeta_j}\left(\left|\hat{\boldsymbol{\alpha}}_j\right|\right)|\boldsymbol{\alpha}_j| - n\sum_{k=1}^{p_2} p'_{\eta_k}\left(\left|\hat{\boldsymbol{\beta}}_k\right|\right)|\boldsymbol{\beta}_k| - n\sum_{l=1}^{p3} p'_{\boldsymbol{\kappa}_l}\left(\left|\hat{\boldsymbol{\gamma}}_l\right|\right)|\boldsymbol{\gamma}_l|$$

The maximum likelihood estimates can be obtained by maximizing the objective function Q, the constant terms in the local approximation of the penalty function are ignored in the above expression as they do not contribute in the estimation. Therefore, the one step sparse estimates $\left(\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{\gamma}}\right)$ may be obtained as follows:

$$\left(\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{\gamma}}\right)=argmin_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}}\left\{\begin{array}{l}\dfrac{1}{2}\left[\left(\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}}\right)',\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)',\left(\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}}\right)'\right][-\Delta^2\ell\left(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}\right)]\begin{bmatrix}\left(\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}}\right)'\\\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\\\left(\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}}\right)'\end{bmatrix}\\+n\sum_{j=1}^{\mathrm{p_1}}\mathrm{p}'_{\zeta_\mathrm{j}}\left(\left|\hat{\boldsymbol{\alpha}}_j\right|\right)|\boldsymbol{\alpha}_\mathrm{j}|+\mathrm{n}\sum_{\mathrm{k}=1}^{\mathrm{p_2}}\mathrm{p}'_{\eta_\mathrm{k}}\left(\left|\hat{\boldsymbol{\beta}}_\mathrm{k}\right|\right)|\boldsymbol{\beta}_k|+n\sum_{l=1}^{p3}p'_{\boldsymbol{\kappa}_l}\left(\left|\hat{\boldsymbol{\gamma}}_l\right|\right)|\boldsymbol{\gamma}_l|\end{array}\right\}$$

Since we are using SCAD penalty, the one step sparse estimator is the one step SCAD and can also be viewed as an adaptive LASSO where the weights are obtained from the SCAD penalty. Since the objective function reduced into a quadratic term plus a weighted L_1 penalty, hence LARS algorithm can be applied to obtain the one step SCAD estimates. Following the similar strategy used by Fan and Li(2001) and Buu et al. (2011), to save the computational cost, we rewrite tuning parameter as

$$\zeta_j=\tau\ SE\left(\hat{\boldsymbol{\alpha}}\right),\eta_k=\tau\ SE\left(\hat{\boldsymbol{\beta}}\right)and\ \kappa_l=\tau\ SE\left(\hat{\boldsymbol{\gamma}}\right)$$

where, $SE\left(\hat{\boldsymbol{\alpha}}\right)$, $SE\left(\hat{\boldsymbol{\beta}}\right)$ and $SE\left(\hat{\boldsymbol{\gamma}}\right)$ are the standard error of maximum likelihood estimates obtained without penalizing the likelihood. In this way, different penalty is used for different regression slope parameters. Tuning parameter τ is selected using the minimum Bayesian information criterion (BIC).

6. SIMULATION

As mentioned by Burton et al. (2006) simulated data should be close to the structure of the real data. Therefore, we simulated the datasets having outcome variable similar to the NHANES cigarette smoking data sets. In the NHANES cigarette smoking data set, we observed higher frequency of either no cigarette smokers (0) or half pack (10) a day and full pack (20) a day cigarette smokers. Also, the non-inflated counts are

observed highly dispersed. Hence, we took three inflated counts, i.e., $\{0,\ 10,\ 20\}$ in the NB distributed response variable and a predictor variable in each model component. The data are simulated from the following MINB model:

$$\begin{cases} \textit{Cumulative logit model}: \begin{cases} \log\left(\dfrac{\Pr(c_1 \le 0}{1-\Pr(c_1 \le 0}\right) = \gamma_{00} + \gamma_1 X_1, \\ \log\left(\dfrac{\Pr(Y \le 10)}{1-\Pr(Y \le 10)}\right) = \gamma_{01} + \gamma_1 X_1, \end{cases} \\ \textit{logit model}: \log\left(\dfrac{\phi}{1-\phi}\right) = \alpha_0 + \alpha_1 X_2, \\ \textit{Negative Binomial } (NB(\mu,\tau))\textit{Model}: \log(\mu) = \beta_0 + \beta_1 X_3;\ \tau \sim \textit{gamma}(\theta,\theta) \end{cases}$$

Data sets are simulated by choosing the population parameters (α_0,α_1)= (-1,2), (β_0,β_1) = (1,1), $(\gamma_{00},\gamma_{01},\gamma_1)$ = (-0.5,0.5,1.2) and θ =3. Each data set includes three covariates (X_1,X_2,X_3) which are independently generated from the uniform distribution. Three sample sizes, small (n = 500), medium (n = 700) and large (n = 900), were chosen keeping in mind the large number of observations (more than 7000) in the survey samples such as NHANES cigarette-smoking data sets (Notice:- NHANES cigarette-smoking data sets for the years :-2005 – 2006, 2007-2008, 2009-2010, 2011-2012 have around 7000 obs.). The sample consists of 10% observations of the cigarette-smoking data set is considered of medium sample size. Using the above population parameters and 1000 simulation runs, firstly, the estimates of the population parameters and their standard errors are obtained for each data set in each run and then the mean of parameter estimates and standard errors along with the standard deviation of the parameter estimates are calculated for all the 1000 runs, which are given in Table 1. As mentioned earlier, the

variance-covariance matrix is approximately estimated by using hessian matrix provided by R Optim() function using finite difference approximation.

Table 1: Average value of the parameter estimates obtained after applying the MINB model on simulated data set.

Sample Size	Parameter	True Value	Estimates		Average SE
			Average	SD	
500	γ_{01}	-0.50	-0.548	0.334	0.276
	γ_{02}	0.50	0.460	0.318	0.267
	γ_1	1.2	1.294	0.608	0.458
	α_0	-1	-1.004	0.212	0.215
	α_1	2	2.013	0.369	0.381
	β_0	1	0.995	0.115	0.123
	β_1	1	1.003	0.182	0.198
	ν	3	3.162	0.696	0.447
700	γ_{01}	-0.50	-0.513	0.289	0.231
	γ_{02}	0.50	0.491	0.279	0.224
	γ_1	1.2	1.232	0.553	0.385
	α_0	-1	-1.002	0.173	0.181
	α_1	2	2.010	0.311	0.322
	β_0	1	0.996	0.099	0.104
	β_1	1	1.002	0.158	0.167
	ν	3	3.111	0.521	0.38
900	γ_{01}	-0.50	-0.516	0.263	0.202
	γ_{02}	0.50	0.489	0.257	0.197
	γ_1	1.2	1.220	0.506	0.338

α_0	-1	-1.009	0.159	0.160
α_1	2	2.015	0.282	0.283
β_0	1	0.998	0.086	0.091
β_1	1	1.004	0.137	0.146
ν	3	3.076	0.466	0.339

In Table 1, we not only found that the average of the parameter estimates is close to the true value but also that the average of SEs of estimates is also close to the standard deviation of the estimates. We also observed that the mean of the standard error for the regression parameter estimates in the loglinear model component, i.e. $\hat{\beta}$, are smaller than the mean of the standard error for the regression parameter estimates in the cumulative logit model component, i.e. $\hat{\gamma}$; this reflects the need of relatively more observations in each category of the cumulative logit model. Our results are in accordance with the results obtained by Su et al. (2013). The 95% CI coverage rate for each sample size is found to be 100%; hence, it is not mentioned explicitly in Table 1.The comparison of the standard error with the standard deviation allows us to evaluate the asymptotic performance of the results hence are included in Table 1.

7. COMPARISON OF THE MINB WITH THE OTHER COUNT MODELS USING SIMULATED DATA

In this section, we compared the MINB model with the other competitive count models frequently used to model the over dispersed counts, mainly the NB regression and ZINB regression models, using the three covariates $\{X_1, X_2, X_3\}$. For this, we simulated a test data set and a training data set with 500 independent observations in each data set. We fitted the count models in the training data set and then applied the fitted

models in the test dataset to get the predicted value of the count outcome. Then the average square loss (ASL) was calculated using the following formula:

$ASL = \frac{1}{n}\sum_{1}^{n}\{\hat{Y}_i - E(Y_i)\}^2$, where, $\hat{Y}$ is vector of n predicted values and Y is the true value.

The averages of ASLs over 1000 runs are given in Table 2, and as expected the MINB model has the smallest ASL. The MINB performs better than the log-linear and ZIP models. This was expected because the simulated datasets are dispersed and violate equidispersion assumption required for the Poisson distribution. We simulated the data to evaluate the performance of the MINB in the situation in which application of the MINB is preferred, i.e., the presence of the over dispersion. However, for the multiple inflated data having equidispersed non-inflated counts, the MIP model is already presented. Moreover, a separate study is needed to observe their (i.e. log-linear, ZIP and MIP models) behavior in the presence of the dispersed counts with multiple inflations.

Specifically, we used all the three covariates in the Poisson and NB model. However, for the ZIP and ZINB models, the variables X_1 and X_2 are taken as covariates for the logit model part whereas X_1 and X_3 are taken as covariates in the loglinear model part. This model specification was adapted by keeping in mind the role of the covariates in the simulation of the data.

Table 2: Comparison of the MINB model with the other models

ASL_{MINB}	$ASL_{LOG-LINEAR}$	ASL_{NB}	ASL_{ZIP}	ASL_{ZINB}
0.26	0.44	0.45	0.50	0.50

8. VARIABLE SELECTION

Finally, the variable selection using the one step SCAD estimator is evaluated as discussed above. We used the MLEs as the initial values for the local approximation of the likelihood and also for the local linear approximation of the penalty function. For variable selection, we simulated a dataset of 500 independent observations by choosing the population parameters $(\alpha_0, \alpha_1, \alpha_2, \alpha_3)$ =(-1,2,0,0), $(\beta_0, \beta_1, \beta_2, \beta_3)$ =(1,1,0,0), $(\gamma_{00}, \gamma_{01}, \gamma_1, \gamma_2, \gamma_3)$=(-0.5,0.5,1.2,0,0) and θ=3. Using these population parameters and 1000 simulation runs, the estimates of the population parameters are obtained for each data set in each run then specificity and sensitivity for each parameter is calculated. The specificity is the percentage at which the zero coefficients are correctly estimated to be zero, and sensitivity is the percentage at which a non-zero coefficient is correctly estimated to be non-zero. These percentages are given in Table 3. The percentage at which each model component is correctly identified (i.e. in each model components, the percentage at which non-zero parameter is estimated as non-zero as well as zero parameters are estimated as zero) is also given in the table.

The possibility of having the confounding variable is also explored in the variable selection process. For this purpose, the same covariate associated with non-zero coefficient present in the cumulative logit model is taken into the logit model component (results are given in Table 4). However, the same covariate associated with non-zero coefficient present in the cumulative logit model or logit mode is also taken into the NB model component and results are given in Tables 5 and 6 respectively. The presence of the same variables in different model parts causes difference in the parameter estimates. However, the performance of the variable selection taking specificity and sensitivity into

the account remain almost similar and does not affect much. The tuning parameters we are taking here are the function of the standard errors (SEs) of the parameter estimates and therefore a greater penalty for larger standard errors is applied. Therefore, the effect of confounding variables in variable selection remains the topic of further research.

However, when all the models have different covariates, one step SCAD not only provides good estimates but also provides the variable selection with high sensitivity. When covariates are different in all the three model components, we used the following model. Table 3 gives the performance of the one step SCAD in such scenario.

$$\begin{cases} \textit{Cumulative logit model}: \begin{cases} \log\left(\dfrac{\Pr(c_1 \le 0}{1-\Pr(c_1 \le 0}\right) = \gamma_{00} + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3, \\ \log\left(\dfrac{\Pr(c_2 \le 10}{1-\Pr(c_2 \le 10}\right) = \gamma_{01} + \gamma_1 X_1 + \gamma_2 X_2, + \gamma_3 X_3, \end{cases} \\ \textit{logit model}: \log\left(\dfrac{\phi}{1-\phi}\right) = \alpha_0 + \alpha_1 X_4 + \alpha_2 X_5 + \alpha_3 X_6, \\ \textit{Negative Binomial}: \log(\mu) = \beta_0 + \beta_1 X_7 + \beta_2 X_8 + \beta_3 X_9 \end{cases}$$

Table 3: Sensitivity and specificity when the covariates associated with the non-zero coefficients are different in all the three model components.
.

Model Component	Sensitivity	Specificity	Correct Selection	Average of non- zero slope parameter estimates (True value)
Cumulative logit	86.6	51.1	43.4	1.049 (1.2)
Logit	100	57.1	57.1	2.032 (2)
Negative Binomial	100	68.2	68.2	1.006 (1)

To demonstrate the confounding effect the same covariate associated with non-zero coefficient are taken in the cumulative logit and the logit model parts. In particular, we used the following model :

$$\begin{cases} \textit{Cumulative logit model}: \begin{cases} \log\left(\dfrac{\Pr(c_1 \le 0}{1-\Pr(c_1 \le 0}\right) = \gamma_{00} + \gamma_1 \boldsymbol{X}_1 + \gamma_2 X_2 + \gamma_3 X_3, \\ \log\left(\dfrac{\Pr(c_2 \le 10}{1-\Pr(c_2 \le 10}\right) = \gamma_{01} + \gamma_1 \boldsymbol{X}_1 + \gamma_2 X_2, + \gamma_3 X_3, \end{cases} \\ \textit{logit model}: \log\left(\dfrac{\phi}{1-\phi}\right) = \alpha_0 + \alpha_1 \boldsymbol{X}_1 + \alpha_2 X_5 + \alpha_3 X_6, \\ \textit{Negative Binomial}: \log(\mu) = \beta_0 + \beta_1 X_7 + \beta_2 X_8 + \beta_3 X_9 \end{cases}$$

Table 4: Sensitivity and specificity when the covariates associated with the non-zero coefficients in the cumulative logit and logit model components are same.

Model Component	Sensitivity	Specificity	Correct Selection	Average of non-zero slope parameter estimates (True value)
Cumulative logit	83.8	49.3	40.2	1.052 (1.2)
Logit	100	58.6	58.6	2.049 (2)
Negative Binomial	100	68.3	68.3	1.011 (1)

In the subsequent discussion, to demonstrate the effect of the same independent variables in the other different model parts the models are given before the corresponding tables.

$$\begin{cases} \textit{Cumulative logit model}: \begin{cases} \log\left(\dfrac{\Pr(c_1 \le 0}{1-\Pr(c_1 \le 0}\right) = \gamma_{00} + \gamma_1 \boldsymbol{X}_1 + \gamma_2 X_2 + \gamma_3 X_3, \\ \log\left(\dfrac{\Pr(c_2 \le 10}{1-\Pr(c_2 \le 10}\right) = \gamma_{01} + \gamma_1 \boldsymbol{X}_1 + \gamma_2 X_2, + \gamma_3 X_3, \end{cases} \\ \textit{logit model}: \log\left(\dfrac{\phi}{1-\phi}\right) = \alpha_0 + \alpha_1 X_4 + \alpha_2 X_5 + \alpha_3 X_6, \\ \textit{Negative Binomial}: \log(\mu) = \beta_0 + \beta_1 \boldsymbol{X_1} + \beta_2 X_8 + \beta_3 X_9 \end{cases}$$

Table 5: Sensitivity and specificity when the covariates associated with the non-zero coefficients in the cumulative logit and NB model components are same.

Model Component	Sensitivity	Specificity	Correct Selection	Average of non-zero slope parameter estimates (True value)
Cumulative logit	88.3	47.7	41.3	1.131 (1.2)
Logit	100	58.5	58.5	2.005 (2)
Negative Binomial	100	65.1	65.1	1.008 (1)

$$\begin{cases} \textit{Cumulative logit model}: \begin{cases} \log\left(\dfrac{\Pr(c_1 \le 0}{1-\Pr(c_1 \le 0}\right) = \gamma_{00} + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3, \\ \log\left(\dfrac{\Pr(c_2 \le 10}{1-\Pr(c_2 \le 10}\right) = \gamma_{01} + \gamma_1 X_1 + \gamma_2 X_2, + \gamma_3 X_3, \end{cases} \\ \textit{logit model}: \log\left(\dfrac{\phi}{1-\phi}\right) = \alpha_0 + \alpha_1 \boldsymbol{X}_4 + \alpha_2 X_5 + \alpha_3 X_6, \\ \textit{Negative Binomial}: \log(\mu) = \beta_0 + \beta_1 \boldsymbol{X}_4 + \beta_2 X_8 + \beta_3 X_9 \end{cases}$$

Table 6: Sensitivity and specificity when the covariates associated with the non-zero coefficients in the logit and NB model components are same.

Model Component	Sensitivity	Specificity	Correct Selection	Average of non- zero slope parameter estimates (True value)
Cumulative logit	80.5	46.1	36.6	0.871 (1.2)
Logit	100	58.6	58.6	2.022 (2)
Negative Binomial	100	67.5	67.5	0.992 (1)

CONCLUSION

We developed the MINB model and applied it on the simulated data sets. The performance of the MINB model was compared with the other competitive count models which are frequently used to model the over dispersed data. The ASL was used to evaluate the performance of the MINB and other models. We found that the MINB performs better than the other models in the data set we simulated. However, an

evaluation of the performance of the model in different scenarios is needed. The variable selection issue was also successfully handled using the one step SCAD variable selection method. Using the simulated data sets to evaluate the performance of the variable selection methods, we found very promising results with very high sensitivity even when any two model components have the same independent variables associated with the non-zero coefficient. The high sensitivity ensures the selection of all the important variables. However, the specificity found to be moderate but variables associated with non-zero coefficient are selected with very high percentage even when two model components have same independent variables associated with the non-zero coefficient. The one step SCAD estimates are found very close to the true value.

REFERENCES

Burton, A., Altman, D.G., Royston, P. and Holder, R.L. (2006). The design of simulation studies in medical statistics. *Statist. Medicine* **25**, 4279–4292.

Buu, A., Johnson, J.N., Li, R. and Tan, X. (2011). New variable selection methods for zero-inflated count data with applications to the substance abuse field. *Statist. Medicine* **30**, 2326-2340.

Centers for Disease Control and Prevention (CDC)[2005 – 2006, 2007-2008, 2009-2010, 2011-2012]. National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. [http://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Questionnaire] Accessed: 11 August 2013.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. Roy. Statist. Soc. Ser. B* **39,** 1-38.

Do, C.B. and Batzoglou, S. (2008). What is the expectation maximization algorithm? *Nat Biotechnol* **26,** 897-899.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Greene, W. H. (1994). Some Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models. Working Paper EC-94-10: Department of Economics, New York University.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34**, 1-14.

NHANES (2005-2012). Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, [2001-2002] [http://www.cdc.gov/nchs/nhanes.htm] Accessed: 11 August 2013.

NHANES (2005-2012). Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination

Survey Questionnaire. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, [2001-2002] [http://www.cdc.gov/nchs/nhanes.htm] Accessed: 11 August 2013.

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Su, X. G., Fan, J., Levine, R., Tan, X., and Tripathi, A. (2013). Multiple-Inflation Poisson Model with ℓ1 Regularization. *Statistica Sinica* **23**, 1071-1090.

Tallis, G. M. and Chesson, P. (1982). Identifability of mixtures. *J. Austral. Math.Soc.Ser. A* **32**, 339-348.

Teicher, H. (1961). Identifiability of mixture. *Ann. Math. Stat.* **32**, 244-248.

Tibshirani, R.J.(1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Zou, H., and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Annals of Statistics* **8**, 1509-1566.

SUPPLEMENTARY MATERIAL

S1 Proof of Proposition 3.1

Taking $\mathfrak{F} = \{\mathbf{F}_c(\mathbf{x}); c \in \mathcal{I}\} \cup \{\mathbf{F}_c(\mathbf{x}); c \notin \mathcal{I}\}$ and noticing that the number of parameters of the MINB depends on c. Taking

$$\mathcal{R}_1^{2\kappa} = \{\tau'_{1c}, \tau'_{2c}, \ldots \tau'_{\kappa c}; c \in \mathcal{I}\} \cup \{\tau_{1c}, \tau_{2c}, \ldots \tau_{\kappa c}; c \notin \mathcal{I}\}$$

Writing

$$H(x) = \omega_0 \cdot NB(\mu_i, \nu) + \omega_1 \cdot \mathbf{I}\{Y = c_1\} + \ldots + \omega_M \cdot \mathbf{I}\{Y = c_M\} = 0, \text{where,}$$

$$\sum_{i=1}^{M} \omega_i = 1 \; \forall Y \notin I^-$$

According to Tallis et al. (1982), the above mixture $\mathcal{H}(\mathbf{x})$ is identifiable if and only if $\mathfrak{F}$ is linearly independent. We are providing the proof of this proposition by contradiction.

In order to prove this proposition, we will also use the definition of linear independence of the set of functions $\mathfrak{F}$. Tallis et al. (1982) mentioned the definition of the linear independence as follows:

A set of functions $\mathfrak{F}$ is said to be linearly independent if for real constant a_i

$\sum_{i=0}^{M} a_i F_i(x) \equiv 0 \Rightarrow a_i = 0$, for $i = 0,1,\cdots M$. More precisely,

$$a_0 \cdot NB(\mu_i, \nu) + a_1 \cdot \mathbf{I}\{Y = c_1\} + \ldots + a_M \cdot \mathbf{I}\{Y = c_M\} \equiv 0 \Rightarrow a_i = 0 \forall i \in \{0,1,\ldots,M\} \ldots 3.1$$

Suppose that $\mathfrak{F}$ is not linearly independent. Therefore, $\sum_{i=0}^{M} a_i F_i(x) \equiv 0 \Rightarrow \exists$ at least one i such that $a_i \neq 0$.

Case-I.

Suppose that $a_i \neq 0$ for $i = 0$. Then by eq. (3.1),

$a_0 \cdot NB(\mu_i, \nu) \equiv 0 \Rightarrow NB(\mu_i, \nu) \equiv 0 \Rightarrow$ Contradiction (as per the model definition of the MINB).

Therefore, a_0=0, this implies that $\mathfrak{F}$ is linearly independent. Therefore, by Tallis et al. (1982) $\mathcal{H}(\mathbf{x})$ is identifiable.

Case-II

Without loss of generality, suppose that $a_i \neq 0$ for any $i = 1, \cdots, M$. Then by eq. (3.1),

$a_i \cdot \mathbf{1}\{Y = c_i\} = 0 \Rightarrow \mathbf{1}\{Y = c_i\} = 0 \Rightarrow c_i \notin \mathcal{I} \Rightarrow$ Contradiction (as per the model definition of the MINB).

Therefore, $a_i = 0, \forall\, i \in \{1, 2, \cdots, M\}$ this implies that $\mathfrak{F}$ is linearly independent. Therefore, by Tallis et al. (1982) $\mathcal{H}(\mathbf{x})$ is identifiable. Therefore, the MINB is identifiable.

S2 Technical Details about the implementation of LARS

As proposed by Zou and Li (2008), implementation of the one step sparse estimator in the MINB model starts with the creation of working data. We are following Buu et al.(2011) but the steps suggested by him are rearranged little bit as per convenience.

Step1:- We find the Cholesky decomposition of

$$\boldsymbol{\Sigma}_0 = \Delta^2 \ell\left(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right)$$

$$\Sigma_0 = L'L$$

using this, we get

$$Y^{*} = L\hat{\boldsymbol{\theta}}$$

where $\hat{\boldsymbol{\theta}}$ is the set of maximum likelihood estimates of the slope parameters for the cumulative logit, logit and NB model respectively.

Now in-order to obtain working data, we proceed as follows:-

Step2:- We fist get the index, i.e.,

$$U = \left\{ j : p'_{\zeta j}\left(\left|\hat{\boldsymbol{\alpha}}_j\right|\right) = 0 \right\} \cup \left\{ k : p'_{\eta_k}\left(\left|\hat{\boldsymbol{\beta}}_k\right|\right) = 0 \right\} \cup \{ l : p'_{\kappa_l}(\left|\hat{\boldsymbol{\gamma}}_l\right| = 0 \} \; and$$

$$V = \left\{ j : p'_{\zeta j}\left(\hat{\boldsymbol{\alpha}}_j\right) \neq 0 \right\} \cup \left\{ k : p'_{\eta k}\left(\left|\hat{\boldsymbol{\beta}}_k\right|\right) \neq 0 \right\} \cup \left\{ l : p'_{\kappa l}\left(\left|\hat{\boldsymbol{\gamma}}_l\right|\right) \neq 0 \right\}$$

$$X_V^{*} = \left\{ x_j^{*} : x_j^{*} = \frac{\tau}{np'_{\zeta j}\left(\left|\hat{\boldsymbol{\alpha}}_j\right|\right)} \ell^{*}{}_j ; \; \forall j \in V \right\} \cup \left\{ x_k^{*} : x_k^{*} = \frac{\tau}{np'_{\eta k}\left(\left|\hat{\boldsymbol{\beta}}_k\right|\right)} \ell^{*}{}_k ; \; \forall k \in V \right\} \cup$$

$$\left\{ x_l^{\{*\}} : x_l^{*} = \frac{\tau}{np'_{\kappa l}\left(\left|\hat{\boldsymbol{\gamma}}_l\right|\right)} \ell_l^{*} ; \; \forall l \in V \right\}$$

where $X^{*} = \left\{ X_U^{*}, X_V^{*} \right\}$

Step 3:- This step involves obtaining the projection matrix in the space of X_U^{*} i.e.

$$H_U = X_U^{*}\left(X^{*'}{}_U X_U^{*} \right)^{-1} X_U^{*}$$

and then computation of

$$Y^{**} = Y^{*} - H_U Y^{*}, X_V^{**} = X_V^{*} - H_U X_V^{*}$$

Step 4:- In this step, we apply LARS algorithm and solve

$$\hat{\boldsymbol{\theta}}_U^{*} = \arg min_{\boldsymbol{\theta}} \left\{ \frac{1}{2}\left| Y^{**} - X^{**}\boldsymbol{\theta} \right| + \tau \sum_{i=1}^{j+k+l} \left|\boldsymbol{\theta}_i\right| \right\}$$

Then, the one step SCAD estimator is given by

$$\boldsymbol{\theta}_u^* \cup \left\{ \boldsymbol{\theta}_j^* : \frac{\tau}{np'_{\tau j}\left(\left|\hat{\boldsymbol{\theta}}_j\right|\right)} \boldsymbol{\theta}_j^* \;\; \forall j \in V \right\}$$

MULTIPLE-INFLATION GENERALIZED POISSON MODEL

by

ARVIND TRIPATHI, KUI ZHANG AND XIAOGANG SU

Submitted in Journal of Royal Statistical Society Series A

Format adapted for dissertation

1. SUMMARY

In modeling count data, the presence of over/under dispersed counts and occurrences of more than expected counts (i.e., inflation in many counts / multiple inflated counts) are two major problems and inappropriate models can lead to substantially misleading results. The available methods, such as the multiple-inflation Poisson model (Su et al., 2013) can accommodate the multiple inflations in count data. However, equidispersion property of the Poisson distribution is not suitable to model the over/under dispersed non-inflated counts. In this paper, we proposed a multiple-inflation generalized Poisson (MIGP) regression model by using a mixture of a cumulative logit model and generalized Poisson model since the generalized Poisson distribution (GPD) (Consul and Jain, 1973) can be used to accommodate the dispersion among count data, the mixing probabilities are formulated with a logistic regression. We applied the MIGP model to simulated data and found that it outperforms other commonly used count models. We also applied the MIGP model to a data from General Social Survey and found results from models without consideration of multiple inflated counts can be misleading.

2. INTRODUCTION

Health surveys provide rich source of publically available data to investigate the association between individual's behaviors and human diseases. Often information about such a behavior is collected in the form of count data which involves dispersed counts. From a publically available data on substance abuse at "National Health and Nutrition Examination Survey (NHANES)" website (NHANES 2011-2012), we noticed the presence of the unusual higher frequency for many counts merely by inspecting the histogram plots (Figures 1 and 2). Here unusual higher frequency means that the counts were in higher frequencies than the expected frequency under the Poisson and/or negative binomial (NB) distribution. Su et al. (2013) referred this as multiple inflated counts. When only count zero is inflated then this is referred as zero inflated counts. However, the inflation in counts may reflect sample-to-sample variation and could just show the variation across the samples. Moreover, it may reflect that some counts in the sample are following different distribution than just a Poisson/ NB distribution. The difference in distribution is mainly caused by the data generating mechanism.

Lambert (1992) observed many zero counts than expected under the Poisson distribution and proposed a zero-inflated Poisson (ZIP) model. Lambert found that the mechanisms leading to the excess zeros consist of two processes with different distribution functions, where one process generates the zero counts while the other process generates the counts following Poisson distribution. She proposed the zero-inflated Poisson (ZIP) regression, with an application to defects in manufacturing.

Figures 1 and 2 clearly show the multiple inflations at 16-20 and at 2-6 for the variables "Age first injected the drug" (0 to 49) and "Last time injected drugs",

respectively (NHANES 2011-2012). The presence of inflated counts may reflect the vulnerable behavior of the certain population at particular age and producing the counts with different mechanism. The "Sexual Behavior" questionnaire survey available at NHANES website is also among the other examples in which we encountered with the multiple inflated counts (NHANES 2011-2012). For example, we found that the count variables such as number of male sex partner (from 0 to 600) in lifetime have a few counts with inflation (Figure 3). The inflation in count 1 may refer monogamous population and may have different counts generating mechanism whereas for the mechanism of inflation in other counts, proper investigation of population is needed.

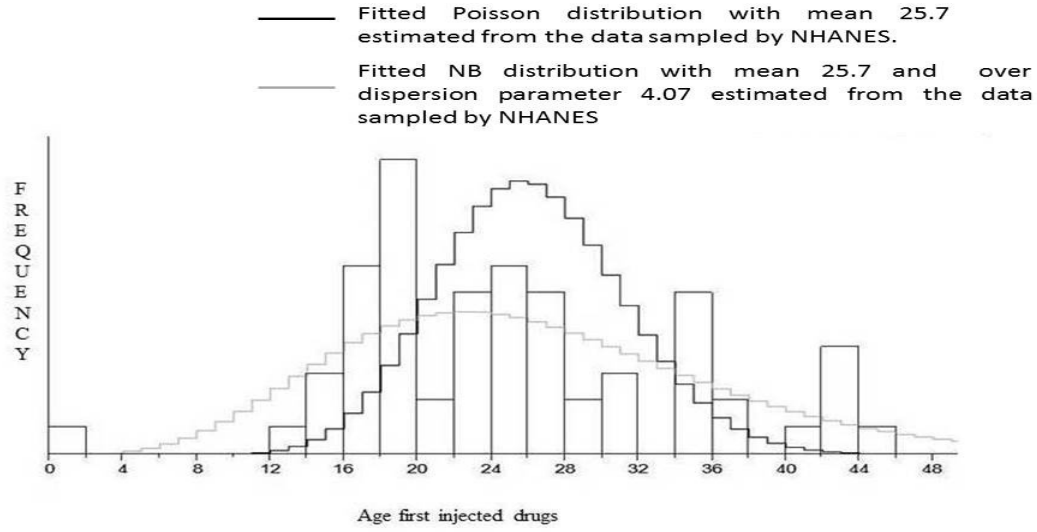


Figure 1. Histogram plot of the variable "Age first injected drugs" along with the fitted Poisson distribution and negative binomial distribution with parameters estimated from the data.

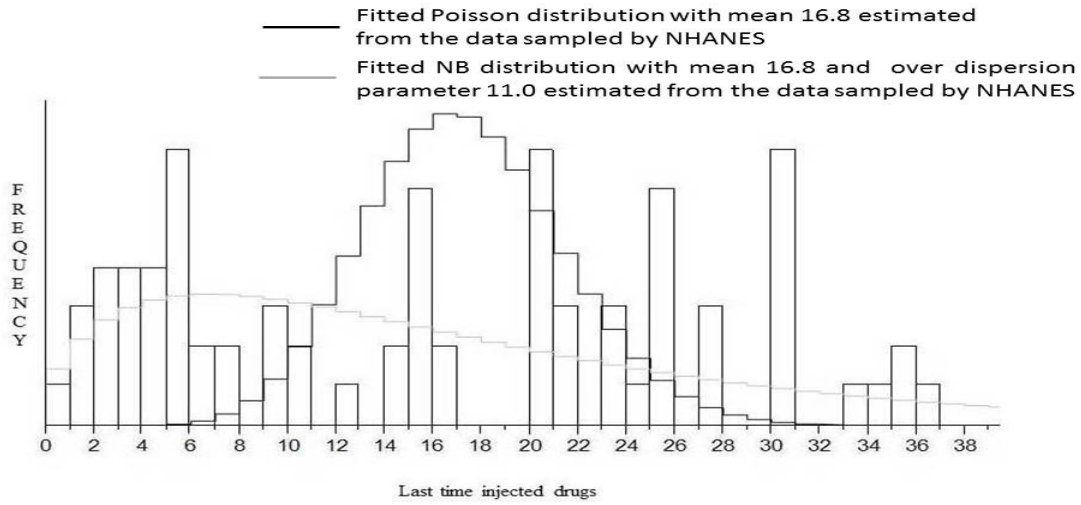


Figure 2. Histogram plot of the variable "Last time injected drugs" along with the fitted Poisson distribution and negative binomial distribution with parameters estimated from the data.

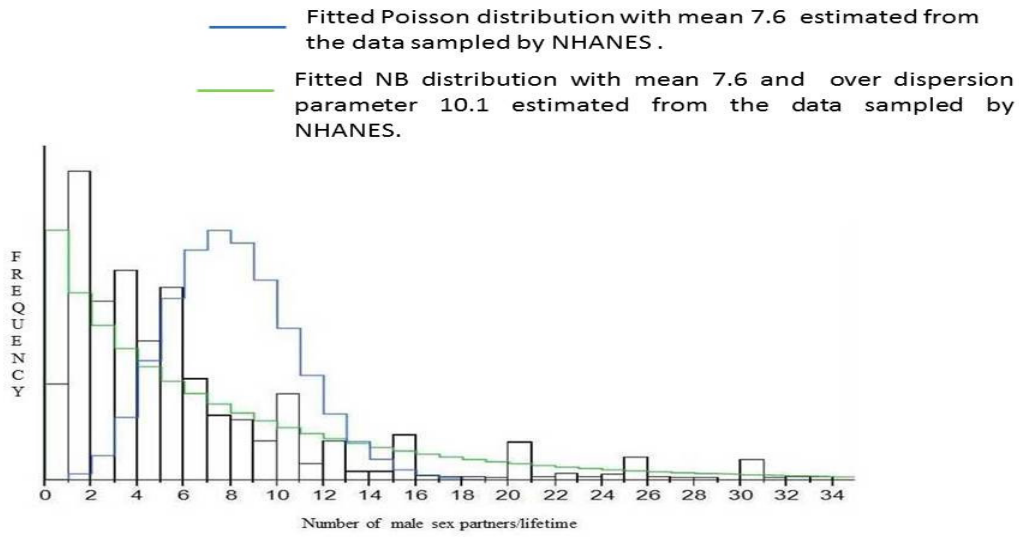


Figure 3. Histogram plot of the variable "number of male sex partners/lifetime" along with the fitted Poisson distribution and negative binomial distribution with parameters estimated from the data.

For data with multiple inflated counts, neglecting such inflations could provide misleading results. Su et al. (2013) proposed a multiple-inflation Poisson (MIP) model for count data with inflations in multiple counts. The MIP model assumes a mixture of a Poisson and a discrete distribution. However, the equidispersion assumption of the Poisson distribution characterized by equality in mean and variance is largely violated in many situations. To deal with this shortcoming of the Poisson distribution, Lagrangian-Poisson distribution, also known as the generalized Poisson distribution (GPD), was proposed by Consul and Jain (1973) and is considered as one of the popular alternative (Johnson et al., 1992). Therefore, when the data exhibit substantial extra (or less)-Poisson variation, or over (or under) dispersion, relative to a Poisson model, the GPD should be preferred over the Poisson distribution. Similar to the NB distribution, the GPD also incorporates a dispersion parameter which follows lognormal distribution instead of a single parameter gamma distribution with the mean 1. Thus, the GPD can be used for modeling both under dispersed as well as over dispersed data (Hilbe, 2011).

To model data with multiple inflations in under or over dispersed counts, we proposed a multiple-inflation generalized Poisson (MIGP) model which would not only be able to provide the adjustment for the multiple inflated counts but will also be able to provide better estimates when non-inflated counts are dispersed. We also provided an EM algorithm and used it along with the numerical optimization to obtain maximum-likelihood estimates of the parameters in the model.

This paper is organized as follows. Section 3 contains detailed formulation of the multiple-inflation generalized Poisson (MIGP) model. The identifiability, and (over and under) dispersion is also discussed in Section 4. The EM algorithm is proposed along

with the numerical optimization to obtain the maximum likelihood (ML) estimates in Section 5. Sections 6 and 7 consist of the application of the MIGP in the simulated data sets to evaluate its inferential performance and its comparison with the frequently used relative existing models that address over dispersion. However, section 8 consists of application of MIGP model in data from General Social Survey (GSS).

3. MULTIPLE-INFLATION GENERALIZED POISSON (MIGP) MODEL

We consider a data set with y_i as the count outcome variable and $\mathbf{X}_i$ as the associated predictor vectors from *n* independent observations namely $\{(y_i, \mathbf{X}_i) : i = 1, \cdots, n\}$. Suppose that the *M* inflated counts are present in the count outcome. Due to the presence of the natural order in the inflated counts, we can arrange them either in the ascending or descending order. Without loss of generality, the inflated counts are arranged in the ascending order and collected in a set *I* such that $I = \{c_1, c_2, \ldots, c_m \ldots, c_M : c_1 \le c_2 \ldots \le c_m \ldots \le c_M\}$. For data with inflated zero counts, we have *M*=1and $I = \{0\}$ .For three inflated counts at 0, 1, and 2, we have *M*=3 then $I = \{0,1,2\}$.

3.1. Model Specification

Analogous to the MIP distribution (Su et al., 2013), the multiple-inflation generalized Poisson (MIGP) model is given as follows:

$$y_i \sim \begin{cases} Discrete\{c_1, c_2, \ldots, c_M ; p_{i1}, p_{i2}, \ldots, p_{iM}\} & \text{with probability}\,(1-\phi_i) \\ GPD(\mu_i, \nu) & \text{with probability}\ \phi_i \end{cases},$$

where the probability density function of the $GPD(\mu, \nu)$ is given as follows:

$$P(Y = y|\nu, \mu) = \mu\left[\mu + (\nu - 1)\, y\right]^{y-1} \nu^{-y} \left(y!\right)^{-1} \exp\left(-\frac{\left[\mu + (\nu - 1)\, y\right]}{\nu}\right)$$

The MIGP model is a mixture of a discrete distribution over the inflated values in I and a generalized Poisson distribution, $GPD(\mu_i, \nu)$, where the mixing probability is specified by ϕ_i.

Furthermore, the parameters p_{im}, ϕ_i and μ_i are further specified by using three different regression models. First, ϕ_i is specified by the logistic regression model, $\log\left(\frac{\phi_i}{1-\phi_i}\right) = \boldsymbol{h}_i^T \boldsymbol{\alpha}$. Second, μ_i is determined by the log-linear regression model, $\log(\mu_i) = \boldsymbol{b}_i^T \boldsymbol{\beta}$. Third, $p_{im}, m = 1, 2, \cdots, M$ are specified through the cumulative logit or proportional odds regression model (as the inflated values in I are naturally ordered). To describe the proportional odds regression model, we first introduce a $(M+1)$-dimensional dummy variable vector $\boldsymbol{\delta}_i = \left\{(\delta_{im}) : \delta_{im} \in R^{(M+1)} \ \forall m = 1, 2, \ldots, M+1\right\}$, where

$$\delta_{im} = \begin{cases} 1_{\{y_i = c_m\}} & \text{for } m = 1, 2, \ldots, M \\ 1_{\{y_i \notin I\}} & \text{for } m = M+1 \end{cases}.$$

Now the cumulative logit or proportional odds regression model is given as follows:

$$\log \frac{\Pr\left(y_i \le c_m | \delta_{i(M+1)} = 0\right)}{\Pr\left(c_m < y_i \le c_M | \delta_{i(M+1)} = 0\right)} = \gamma_{0m} + \boldsymbol{g}_i^T \boldsymbol{\gamma}_1, \forall m = 1, 2, \ldots, M-1.$$

where $\boldsymbol{\gamma} = \left(\gamma_{01}, \gamma_{02}, \ldots, \gamma_{0(M-1)}, \boldsymbol{\gamma}_1^T\right)^T$ is a vector for the regression parameters involved in the cumulative logit model and includes the $M-1$ intercept parameters, i.e., γ_{0j}'s and a slope parameter vector $\boldsymbol{\gamma}_1^T$. Moreover, $\{\boldsymbol{h}_i, \boldsymbol{b}_i, \boldsymbol{g}_i\}$ and $\{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}\}$ represent the covariate vectors of the appropriate dimensions (consist of the selected components from the $\mathbf{X}_i$) and corresponding regression parameter vectors respectively.

The above model specifications can be further written as follows:

$$\mu_i = \exp\left(\boldsymbol{b}_i^T \boldsymbol{\beta}\right) \text{.....(1.1)}$$

$$\phi_i = \text{expit}\left(\boldsymbol{h}_i^T \boldsymbol{\alpha}\right) \text{ .....(1.2)}$$

and

$$p_{im} = P\left(y_i = c_m \mid \delta_{i(M+1)} = 0\right)$$

$$= \begin{cases} \text{expit}\left(\gamma_{01} + \boldsymbol{g}_i^T \boldsymbol{\gamma}_1\right) & for\ m = 1 \\ \text{expit}\left(\gamma_{0m} + \boldsymbol{g}_i^T \boldsymbol{\gamma}_1\right) - \text{expit}\left(\gamma_{0(m-1)} + \boldsymbol{g}_i^T \boldsymbol{\gamma}_1\right) & for\ m = 2,3,\ldots,M-1 \\ 1 - \text{expit}\left(\gamma_{0(M-1)} + \boldsymbol{g}_i^T \boldsymbol{\gamma}_1\right) & for\ m = M \end{cases} \text{.....1.3}$$

where the function $\text{expit}(x)$ is given as $\text{expit}(x) = e^x \Big/ \left(1 + e^x\right)$.

3.2. Dispersion

Different from the dispersion present in the distribution of non-inflated counts, the multiple-inflation (MI) counts themselves induce over and under dispersion. As noted by Su et al. (2013), the inflation present in the multiple counts causes the over dispersion or under dispersion. To illustrate, the expected value and variance of y_i are first obtained as follows:

$$E\left(y_i\right) = E(E\left(y_i | \boldsymbol{\delta}_i\right) = \left(1 - \phi_i\right) \sum_{m=1}^{M} c_m p_{im} + \phi_i \exp\left(\boldsymbol{b}_i^T \boldsymbol{\beta}\right)$$

and

$$Var\left(y_i\right) = E\left(y_i\right) + \left(1 - \phi_i\right) \left[\sum_{m=1}^{M} c_m^2 p_{im} - \left(1 - \phi_i\right) \left(\sum_{m=1}^{M} c_m p_{im} \right)^2 - \sum_{m=1}^{M} c_m p_{im} \right]$$

$$+ \phi_i \exp\left(\boldsymbol{b}_i^T \boldsymbol{\beta}\right) \left(\nu^2 + \exp\left(\boldsymbol{b}_i^T \boldsymbol{\beta}\right) - \phi_i \exp\left(\boldsymbol{b}_i^T \boldsymbol{\beta}\right) - 1 \right)$$

$$\Rightarrow Var\left(y_i\right) = E\left(y_i\right) + \xi_i$$

where the additive term ξ_i is given as follows:

$$\xi_i = \left(1 - \phi_i\right) \left[\sum_{m=1}^{M} c_m^2 p_{im} - \left(1 - \phi_i\right) \left(\sum_{m=1}^{M} c_m p_{im} \right)^2 - \sum_{m=1}^{M} c_m p_{im} \right]$$

$$+ \phi_i \exp\left(\boldsymbol{b}_i^T \boldsymbol{\beta}\right) \left(\nu^2 + \exp\left(\boldsymbol{b}_i^T \boldsymbol{\beta}\right) - \phi_i \exp\left(\boldsymbol{b}_i^T \boldsymbol{\beta}\right) - 1 \right)$$

The presence of the additive term ξ_i suggests the possible over or under dispersion. The positive value of ξ_i gives the over dispersion. Therefore, the condition for the over dispersion (i.e., $\xi_i > 0$) is given as follows:

$$\phi_i \exp\left(\boldsymbol{b}_i^T \boldsymbol{\beta}\right)\left[\nu^2 + (1-\phi_i)\exp\left(\boldsymbol{b}_i^T \boldsymbol{\beta}\right) - 1\right] > [(1-\phi_i)\{\sum_{m=1}^{M} c_m (1-c_m) + (1-\phi_i)\left(\sum_{m=1}^{M} c_m p_{im}\right)^2\}]$$

Also the condition for the under dispersion is given as follows:

$$\phi_i \exp\left(\boldsymbol{b}_i^T \boldsymbol{\beta}\right)\left[\nu^2 + (1-\phi_i)\exp\left(\boldsymbol{b}_i^T \boldsymbol{\beta}\right) - 1\right] < [(1-\phi_i)\{\sum_{m=1}^{M} c_m (1-c_m) + (1-\phi_i)\left(\sum_{m=1}^{M} c_m p_{im}\right)^2\}]$$

The left side of the inequality involves the parameters of the generalized Poisson model whereas the right part of the inequality involves the parameters of the discrete distribution. If no inflation in the counts is present (i.e., $c_j = 0$) and the counts are coming only from the GPD (i.e., $\phi_i = 1$), the above expression reduces into the GPD and provides the over dispersion for the $\nu > 1$ and the under dispersion for the $\nu < 1$. As we mentioned that the MIGP also induces the under dispersion in the some cases, e.g., for the two inflated counts 0 and 1 with the probability 0.25 and 0.75, and for the fixed value of $\phi, \mu \text{ and } \nu$ such as 0.7, 2 and 0.03, we get the under dispersion. However after changing the ϕ_i to 0.50, we get the over dispersed counts as per the given inequalities (For the derivation of $Var(y_i)$ see Supplementary Material).

4. IDENTIFIABILITY

In order to provide the valid inference for a model, the model must be identifiable which means two distinct sets of parameters are not giving the same distribution. However, the identifiability could also be achieved for a model by imposing certain set of the restrictions known as identification conditions. A model can be identifiable, partially

identifiable or non-identifiable. In particular, even if a model is non-identifiable, but the true values of a certain subset of the model parameters could be found then the model is called partially identifiable.

According to Tallis et al. (1982), the κ' dimensional parameter space $\mathcal{R}_1^{\kappa'}$ with the restriction that it consist of the finite number of the elements, i.e., $\mathcal{R}_1^{\kappa'} = \{\alpha_1, \alpha_2, ...\alpha_{\kappa'}\}$, the mixture of the distribution $\{\mathbf{F}_1(\mathbf{x}), \mathbf{F}_2(\mathbf{x}), \ldots, \mathbf{F}_\mathbf{i}(\mathbf{x}), \ldots, \mathbf{F}_\eta(\mathbf{x})\}$ *i.e.* $\mathbf{H}(x)$ can be written as $\mathcal{H}(\mathbf{x}) = \sum_{i=1}^{\eta} \omega_i \mathbf{F}_i(\mathbf{x})$ where $\sum_{i=1}^{\eta} \omega_i = 1$. The authors found that the above mixture is identifiable if and only if $H(\mathbf{x}) = \{\mathbf{F}_1(\mathbf{x}), \mathbf{F}_2(\mathbf{x}), \ldots, \mathbf{F}_\mathbf{i}(\mathbf{x}), \ldots, \mathbf{F}_\eta(\mathbf{x})\}$ is linearly independent.

Proposition 4.1

The MIGP model is identifiable (See Supplementary Material).

5. MAXIMUM LIKELIHOOD ESTIMATION

The three regression models mentioned by the equations 1.1 to 1.3 fully specify the MIGP model. Utilizing these, the likelihood function is given by

$$\pi_{ic} = \Pr\{y_i = c\} = \begin{cases} (1-\phi_i)\, p_{im} + \phi_i \cdot \Psi_i & \text{for } c \in I \text{ and } m = 1, 2, \ldots, M \\ \phi_i \cdot \Psi_i & \text{for } c \notin I \end{cases}$$

where $\Psi_i \sim GPD(\mu_i, \nu)$ and is given by:

$$\Psi_i = \mu_i \left[\mu_i + (\nu - 1) y_i\right]^{y_i - 1} \nu^{-y_i} (y_i!)^{-1} \exp\left(-\frac{\left[\mu_i + (\nu - 1) y_i\right]}{\nu}\right)$$

The likelihood function $\ell(\boldsymbol{\Theta})$ (where $\boldsymbol{\Theta} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \nu\}$) of the MIGP model is given by:

$$\ell(\boldsymbol{\Theta}) = \prod_{i=1}^{n} \left[\prod_{m=1}^{M} \left\{(1-\phi_i)\, p_{im} + \phi_i \Psi_i\right\}^{\delta_{im}} \right] \cdot (\phi_i \Psi_i)^{\delta_{i(M+1)}}$$

The corresponding log-likelihood $L(\theta)$ is given as follows:

$$L(\Theta) = \log \ell(\Theta) =$$

$$\sum_{i=1}^{n} L_i(\Theta) = \sum_{i=1}^{n} \left\{ \left[\sum_{m=1}^{M} \delta_{im} \log\left\{ (1-\phi_i)\, p_{im} + \phi_i \Psi_i \right\} \right] + \delta_{i(M+1)} \log\left(\phi_i \Psi_i\right) \right\}$$

Su et al. (2013) used EM algorithm and BFGS quasi-Newton method to find the estimates of the MIP model and also recommended the use of limited-memory variant, L-BFGS when the number of parameters is very large. Here we use the EM algorithm and an optimization algorithm- Nelder Mead simplex algorithm implemented in R function optim() for numerical optimization.

5.1. EM Algorithm

Dempster et al. (1977) proposed the expectation-maximization (EM) algorithm to find the maximum likelihood estimates of the parameters of the mixture models. This algorithm consists of the iterations of the two steps namely E-Step and M-step. The E-step involves the computation of the conditional expectation of the complete-data log likelihood conditional upon the observed data and the current parameter estimates. In mixture models, the complete data generally consists of a latent variable $\boldsymbol{Z}$ and the observed data X. The M-step involves the maximization of the expectation of the log likelihood function, in the every iteration. Following this, we first introduce the latent random variables Z_{im} as follows:

$$Z_{im} = \begin{cases} 1 & \text{if } y_i = c_{im}, \forall c_{im} \in I : m = 1, 2, \ldots, M \text{ and } i \in I^{+} \\ 0 & \text{if } y_i \in \text{GPD} \end{cases}$$

Subscripts i and m in Z_{im} represent the i^{th} random variable for the m^{th} inflated count. Similarly, for the $(M+1)^{\text{th}}$ count, the i^{th} random variable $Z_{i(M+1)}$ is given by

$$Z_{i(M+1)} = \begin{cases} 1 & \text{if } y_i \in \text{GPD} \\ 0 & \text{otherwise} \end{cases} \quad \text{and also } Z_{i(M+1)} = 1 - \sum_{i=1}^{m} Z_{im}.$$

The random variables Z_{im} for the given values of m and the random variable $Z_{i(M+1)}$ are partially observable. Indeed, the value of the variable Z_{im} is only known when $y_i \neq c_m$ and it is unknown otherwise.

Finally, the EM algorithm reduces into the estimation of the $\boldsymbol{\alpha}$ by using the logistic regression model and by taking the $Z_{i(M+1)}$ as the response variable, the estimation of the $\boldsymbol{\beta}$ and $\boldsymbol{\nu}$ by using the generalized Poisson regression model with y_i as the response variable and the $Z_{i(M+1)}$ as a weight, and also the estimation of the $\boldsymbol{\gamma}$ using a form similar to the proportional odds model.

After including the observed and the latent data, the complete data is given as $Y_c = \{Y_i, Z_{im}\}$. Furthermore, the probability density function of the complete data is given as follows:

$$f(Y_c \mid \boldsymbol{\Theta}) = f(Y_i, Z_{im} \mid \boldsymbol{\Theta}),$$

where Θ is the associated parameter space, i.e., $\boldsymbol{\Theta} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \nu\}$. This can be further written as:

$$f\left(Y_i, Z_{im} \mid \boldsymbol{\Theta}\right) = f(Y_i \mid Z_{im}, \boldsymbol{\Theta}) f(Z_{im} \mid \boldsymbol{\Theta})$$

Hence, complete data likelihood can be written as follows:

$$\ell = \prod_{i=1}^{n}\prod_{m=1}^{M} f\left(Y_i, Z_{im} \mid \boldsymbol{\Theta}\right)$$

$$\prod_{i=1}^{n}\prod_{m=1}^{M}\left\{f\left(Y_i, Z_{im} = 1 \mid \boldsymbol{\Theta}\right)\right\}^{Z_{im}}\left\{f\left(Y_i, Z_{im} = 0 \mid \boldsymbol{\Theta}\right)\right\}^{Z_{i(M+1)}}$$

or

$$= \prod_{i=1}^{n}\prod_{m=1}^{M}\left\{f\left(Y_i \mid Z_{im} = 1, \boldsymbol{\Theta}\right) f\left(Z_{im} = 1 \mid \boldsymbol{\Theta}\right)\right\}^{Z_{im}}\left\{f\left(Y_i \mid Z_{im} = 0, \boldsymbol{\Theta}\right) f\left(Z_{im} = 0 \mid \boldsymbol{\Theta}\right)\right\}^{Z_{i(M+1)}}$$

$$\Rightarrow \ell = \prod_{i=1}^{n}\prod_{m=1}^{M}\left\{p_{im}\left(1-\phi_i\right)\right\}^{Z_{im}}\left\{\Psi_i\phi_i\right\}^{Z_{i(M+1)}}$$

Therefore, the log-likelihood of complete data is given by:

$$L_c\left(\Theta \mid Y_c\right) = \sum_{i=1}^{n}\left[\sum_{m=1}^{M}\left\{Z_{im}\left(\log p_{im}\right) + Z_{im}\log\left(1-\phi_i\right)\right\} + Z_{i(M+1)}\left(\log\phi_i + \log\Psi_i\right)\right]$$

However,

$$\sum_{m=1}^{M} Z_{im}\log\left(1-\phi_i\right) + Z_{i(M+1)}\log\left(\phi_i\right) = Z_{i(M+1)}\boldsymbol{h}_i^T\boldsymbol{\alpha} - \log\left(1+\exp\left(\boldsymbol{h}_i^T\boldsymbol{\alpha}\right)\right)$$

Therefore, we get,

$$L_c\left(\boldsymbol{\Theta} \mid Y_c\right) = \sum_{i=1}^{n}[\{\sum_{m=1}^{M} Z_{im}\left(\log p_{im}\right)\} - \log\left(1+\exp\left(\boldsymbol{h}_i^T\boldsymbol{\alpha}\right)\right) + Z_{i(M+1)}\left(\boldsymbol{h}_i^T\boldsymbol{\alpha} + \log\Psi_i\right)]$$

where

$$\Psi_i = \mu_i\left[\mu_i + \left(\nu-1\right)y_i\right]^{y_i-1}\nu^{-y_i}\left(y_i!\right)^{-1}\exp\left(-\frac{\left[\mu_i + \left(\nu-1\right)y_i\right]}{\nu}\right) \; s.t.\log\left(\mu_i\right) = \boldsymbol{b}_i^T\boldsymbol{\beta}$$

This reduces into

$$\sum_{i=1}^{n}[\{\sum_{m=1}^{M} Z_{im}\left(\log p_{im}\right)\} + Z_{i(M+1)}\left(\log\Psi_i\right) + Z_{i(M+1)}h_i^T\boldsymbol{\alpha} - \log\left(1+\exp\left(h_i^T\boldsymbol{\alpha}\right)\right)]$$

Now since $L_c\left(\boldsymbol{\Theta}|Y_c\right)$ is linear in the Z_{im} for each $m = 1, 2, \cdots, M$ and $Z_{i(M+1)}$, therefore, the $(k+1)^{\text{th}}$ iteration of the algorithm involves there placement of the Z_{im} for $m = 1, 2, \cdots, M$ and $Z_{i(M+1)}$ with their conditional expectations, conditioned on observed data y_i and parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ and ν. However, in more details, the EM algorithm at the $(k+1)^{\text{th}}$ iteration involves the following steps:

The E step involves the computation of the conditional expectation of Z_{im} for every $m = 1, 2, \cdots, M$ and $Z_{i(M+1)}$ denoted by $\widehat{Z}_{im}^{(k)}$ and $\widehat{Z}_{i(M+1)}^{(k)}$ (i.e.,

$E\left[Z_{im} | Y_i, \boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\gamma}^{(k)}, \nu^{(k)} \right] = \hat{Z}_{im}^{(k)}$ and $E\left[Z_{i(M+1)} | Y_i, \boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\gamma}^{(k)}, \nu^{(k)} \right] = \hat{Z}_{i(M+1)}^{(k)}$)

respectively,

$$\begin{cases} \hat{Z}_{im}^{(k)} = \dfrac{\left(1-\left(\phi_i \mid \hat{\boldsymbol{\alpha}}^{(k)}\right)\right) p_{im} \delta_{im}}{\left(1-\left(\phi_i \mid \hat{\boldsymbol{\alpha}}^{(k)}\right)\right) p_{im} + \left(\phi_i \mid \hat{\boldsymbol{\alpha}}^{(k)}\right)(\psi_i \mid \hat{\boldsymbol{\beta}}^{(k)}, \hat{\nu}^{(k)})} \text{ if } y_i \in I \text{ and } m = 1, 2 \ldots, M \\ \hat{Z}_{i(M+1)}^{(k)} = 1 - \sum_{m=1}^{M} \hat{Z}_{im}^{(k)} \end{cases}$$

So that,

$$Q\left(\boldsymbol{\Theta} \mid \hat{\boldsymbol{\Theta}}^{(k)}\right) = E\left(L_c\left(\boldsymbol{\Theta} \mid Y_c\right) | Y, \hat{\boldsymbol{\Theta}}^{(k)}\right) = \sum_{i=1}^{n} Q_{1i}(\boldsymbol{\gamma} \mid \hat{\boldsymbol{\Theta}}^{(k)}) + \sum_{i=1}^{n} Q_{2i}(\boldsymbol{\beta}, \nu \mid \hat{\boldsymbol{\Theta}}^{(k)}) + \sum_{i=1}^{n} Q_{3i}(\boldsymbol{\alpha} \mid \hat{\boldsymbol{\Theta}}^{(k)})$$

where

$$Q_{1i}\left(\gamma \mid \hat{\boldsymbol{\Theta}}^{(k)}\right) = \sum_{m=1}^{M} \hat{Z}_{im}^{(k)} \left(\log p_{im}\right),$$

$$Q_{2i}\left(\beta, \nu \mid \hat{\boldsymbol{\Theta}}^{k}\right) = \hat{Z}_{i(M+1)}^{(k)} \left(\log \Psi_i\right),$$

$$= \hat{Z}_{i(M+1)}^{(k)} \log\left\{ \left[\exp\left(\boldsymbol{b}_i^T \boldsymbol{\beta}\right) + (\nu - 1) y_i\right]^{y_i - 1} \nu^{-y_i} \left(y_i !\right)^{-1} \exp\left(-\frac{\left[\exp\left(\boldsymbol{b}_i^T \boldsymbol{\beta}\right) + (\nu - 1) y_i\right]}{\nu} \right) \right\}$$

$$\text{and } Q_{3i}\left(\boldsymbol{\alpha} \mid \hat{\boldsymbol{\Theta}}^{(k)}\right) = \hat{Z}_{i(M+1)}^{(k)} h_i^T \boldsymbol{\alpha} - \log(1 + \exp\left(h_i^T \boldsymbol{\alpha}\right))$$

In the *M*-step, we estimate $\boldsymbol{\beta}$ *and* ν by the maximization of $Q_{2i}\left(\boldsymbol{\beta}, \nu \mid \hat{\boldsymbol{\Theta}}^{k}\right)$ which is the log-likelihood for a weighted generalized Poisson regression of y_i on $\boldsymbol{b}_i$ with the weights $\hat{Z}_{i(M+1)}^{(k)}$. In the *M*-step, we also estimate $\boldsymbol{\gamma}$ and this step involves the maximization of $Q_{1i}\left(\boldsymbol{\gamma} \mid \hat{\boldsymbol{\Theta}}^{(k)}\right)$ which has a form similar to the log-likelihood obtained from a proportional odds model with responses $\hat{Z}_{im}^{(k)}$ .The conjugate gradient method

which does not store a matrix is used for the fast and straight forward maximization in this step. However to get the estimates of $\boldsymbol{\alpha}$, in the *M*-step, we maximize $Q_{3i}\left(\boldsymbol{\alpha} \mid \widehat{\boldsymbol{\Theta}}^{(k)}\right)$ which has a form similar to the log-likelihood obtained from an un-weighted binomial logistic regression of $\widehat{Z}_{i(M+1)}^{(k)}$ on $\boldsymbol{h}_i$.

A similar strategy used by Su et al. (2013) is employed to supply the initial values for the parameters. Specifically, a truncated Poisson model is fit to the data with no inflated counts and a cumulative logit model to the inflated counts. The estimates obtained after fitting these models are used as the initial values for the $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ respectively. The estimate of the index of dispersion is used as the initial value for the ν. The logistic regression model is used to obtain the initial value for the $\boldsymbol{\alpha}$, for this a dichotomous variable with 0 for the inflated and 1 for the non-inflated counts is created and used. However, these values are not directly supplied as the initial values to the EM algorithm, rather these are used as the initial values in a numerical optimization of the likelihood first and a second set of the estimates is obtained which is used as the initial values for the EM algorithm to aid convergence. Specifically, a three step procedure to obtain the maximum likelihood estimates for the MIGP model is used as follows:

(1) As we discussed, the regression parameter estimates of the truncated Poisson model, cumulative logit model and logit model along with the index of dispersion are obtained.

(2) The estimates obtained in the previous step are used as the initial values for the numerical optimization of the log likelihood (for this purpose the default method present in R optim(), i.e., a derivative-free optimization routine, Nelder-Mead simplex algorithm is used) and another set of estimates is obtained

(3) The estimates obtained in the second step are used as the initial values in the EM algorithm and the maximum likelihood estimates are obtained by updating the estimates in EM algorithm until convergence.

Do et al. (2008) has mentioned that in most of the non-concave optimization methods, the EM algorithm provides assurance only for convergence to a local optimum of the objective function. Therefore, the three steps procedure is used to facilitate the convergence of the EM algorithm to the global optimum. In addition, the Hessian matrix obtained from the R function optim() is used to approximately estimate the variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ (*i.e.* $\boldsymbol{\Sigma} = \text{Cov}(\hat{\boldsymbol{\theta}})$) via observed Fisher's information matrix by $\widehat{\boldsymbol{\Sigma}} = \{-\boldsymbol{H}\}^{-1}$. The optim() function uses the finite difference method to approximate hessian matrix.

6. SIMULATION

We simulated data sets that are close to the real data in structure (Burton et al., 2006) and had an outcome variable that are similar in inflated counts to the NHANES cigarette smoking data sets. In the NHANES data, high frequency of the no cigarette smokers (0), half pack (10) a day and full pack (20) a day cigarette smokers is observed along with the highly dispersed counts of the smokers who smoke other than that many (i.e., 0, 10 and 20) cigarettes a day. Therefore, we took three counts, i.e., {0, 10, 20} as the inflated counts to form the following MIGP model:

$$
\begin{cases}
\text{Cumulative logit model}: \begin{cases} \log\left(\dfrac{\Pr(c_1 \le 0)}{1-\Pr(c_1 \le 0)}\right) = \gamma_{00} + \gamma_1 X_1, \\ \log\left(\dfrac{\Pr(c_2 \le 10)}{1-\Pr(c_2 \le 10)}\right) = \gamma_{01} + \gamma_1 X_1, \end{cases} \\
\text{logistic}: \log\left(\dfrac{\phi}{1-\phi}\right) = \alpha_0 + \alpha_1 X_2, \\
\text{Generalised Poisson}: \log(\mu) = \beta_0 + \beta_1 X_3 \ \text{ and} \\
f(Y = y | X_1, \beta, \nu) = \left[\mu + (\nu - 1) y\right]^{y-1} \nu^{-y} (y!)^{-1} \exp\left(-\dfrac{\left[\mu + (\nu - 1) y\right]}{\nu}\right)
\end{cases}
$$

The three covariates (X_1, X_2, X_3) were independently generated from the uniform distributions. The following set of the parameters were used to simulate the data sets (α_0, α_1) =(-1, 2), (β_0, β_1) = (1, 1), $(\gamma_{00}, \gamma_{01}, \gamma_1)$ =(-3, 0.5, 0.5) and ν = 3.Three sample sizes small (n = 500), medium (n = 700) and large(n = 900) were chosen due to the presence of the large number of the observations in the survey samples such as NHANES data sets. We used the population parameters and covariates to generate an outcome variable following a GPD with 0, 10 and 20 inflated counts. 1000 datasets were simulated.

The MIGP model was then applied to each data set to obtain the estimates of the parameters. The estimates of the population parameters and their standard errors were obtained for each data set in each run. The mean of the parameter estimates and the standard errors along with the standard deviation of the parameter estimates were calculated for all the 1000 simulated data sets, which are given in Table 1. The variance-covariance matrix was approximately estimated by using the hessian matrix provided by R Optim() function that uses finite difference approximation.

Table 1: Average value of the parameter estimates obtained after applying the MIGP model on simulated data sets.

Sample Size	Parameter	True Value	Estimates		Average SE
			Average	SD	
500	γ_{01}	-3	-2.124	0.231	0.754
	γ_{02}	0.5	0.782	0.196	0.283
	γ_1	0.5	0.298	0.384	0.492
	α_0	-1	-1.167	0.009	0.230
	α_1	2	2.098	0.003	0.371
	β_0	1	1.027	0.005	0.172
	β_1	1	0.946	0.003	0.244
	ν	3	2.954	0.010	0.281
700	γ_{01}	-3	-2.122	0.180	0.674
	γ_{02}	0.5	0.781	0.153	0.240
	γ_1	0.5	0.299	0.299	0.416
	α_0	-1	-1.167	0.007	0.196
	α_1	2	2.098	0.002	0.313
	β_0	1	1.027	0.004	0.146
	β_1	1	0.946	0.002	0.206
	ν	3	2.954	0.008	0.239
900	γ_{01}	-3	-2.118	0.166	0.571
	γ_{02}	0.5	0.783	0.141	0.210
	γ_1	0.5	0.297	0.277	0.365
	α_0	-1	-1.168	0.006	0.172
	α_1	2	2.098	0.002	0.275

β_0	1	1.028	0.004	0.128
β_1	1	0.946	0.002	0.181
ν	3	2.953	0.007	0.209

From Table 1, we can see that the average of the parameter estimates is close to their true values. The standard deviations are smaller than the SEs. This might be because the computation of the standard errors is based on the hessian matrix obtained in the numerical optimization. However, the standard deviation was computed for the parameter estimates obtained after the application of the EM algorithm. Not much difference in the mean of the estimates is found across the different sample size. The difference in the mean estimates for the cumulative logit model from their true values may be due to the difference in the frequency of the inflated counts across three levels. Most of the data sets were considered more likely to have higher frequency for individuals smoking 10 cigarettes a day during simulation. The 95% CI coverage rate for each sample size is found to be 100% hence it is not mentioned explicitly in Table 1.The comparison of the standard error with the standard deviation allows us to evaluate the asymptotic performance of the results hence are included in Table 1.

7. COMPARISON OF MIGP WITH OTHER RELATED COUNT MODELS USING SIMULATED DATA

We compared the MIGP model with the other related count models, i.e., the NB and ZINB models along with loglinear and ZIP models using the three covariates (X_1, X_2, X_3).

Specifically, we used all the three covariates in the Poisson and NB model. However, for the ZIP and ZINB models, the variables X_1 and X_2 are taken as covariates

for the logit model part whereas X_1 and X_3 are taken as covariates in the loglinear model part. This model specification was adapted by keeping in mind the role of the covariates in the simulation of the data.

For the comparison, the test and the training data sets with 500 independent observations are simulated. We fitted the models in the training dataset and then applied the fitted models in the test dataset to get the predicted value of the count outcome. Then the mean average square loss (ASL) over 1000 data sets was calculated and given in Table 2, where $\text{ASL} = \frac{1}{n}\sum_{1}^{n}\{\hat{Y}_i - E(Y_i)\}^2$ and $\hat{Y}$ is vector of *n* predicted values and *Y* is the true value.

The MIGP performs better than the log-linear and ZIP models (Table 2) with a smallest average ASL of 0.163 while the log-linear model and the ZIP model have an average ASL of 0.511 and 0.996, respectively. This was expected because the simulated datasets are dispersed and violate equidispersion assumption required for the Poisson distribution. We simulated the data to evaluate the performance of the MIGP in the situation in which application of the MIGP is preferred, i.e., presence of over/under dispersion. However, for the data having equidispersed non-inflated counts the MIP model is already available. Moreover, a separate study is needed to observe their (i.e., log-linear, ZIP and MIP models) behavior in the presence of dispersed counts with multiple inflations. The model, zero-inflated generalized Poisson was not used for comparison due to the lack of the appropriate R package to model the ZIGP.

Table 2: Comparison of the MIGP model with the other models.

ASL_{MIGP}	$ASL_{LOG-LINEAR}$	ASL_{NB}	ASL_{ZIP}	ASL_{ZINB}
0.163	0.511	0.547	0.996	1.004

8. REAL DATA ANALYSIS

We provided several examples to bring forth the fact about the presence of the multiple inflated counts in the dispersed count settings in the social science studies. However, this fact has been mostly overlooked till date and the analysis were performed without incorporating multiple inflated counts. In this section, we applied the MIGP model to a data from General Social Survey (GSS, 2012).

Here the "GSS 2010 merged" (GSS, 2012) data has 4,901 observations and 1,223 variables. We only considered the complete cases of 1,784 subjects. We used "How MANY SEX PARTNERS R HAD IN LAST 5 YEARS" as the response variable and sexual orientation, general happiness, and change in financial situation as the predictors. The variable sexual orientation has three levels, i.e., homosexual (gay, lesbian), bisexual and heterosexual (or straight). The three levels very happy, pretty happy and not too happy were also used for general happiness whereas information about change in financial situation was gathered as better, worse and stayed same. In the following discussion, we took 5% as the level of significance.

Without considering any inflation and after applying the Poisson regression model, we found strong association of number of sex partners in last 5 years with sexual orientation (p-value < 0.0001), financial situation (p-value < 0.0001) and happiness (p-value = 0.002). Furthermore, after using the NB regression model, we obtained the

similar association between sexual orientation (p-value < 0.005), financial situation (p-value < 0.0008), and happiness (P=0.03) with the response variable. However, after applying the ZIP model, no variable was found significant in zero model while sexual orientation (p-value < 0.0001), financial situation (p-value < 0.0001) and happiness (p-value = 0.003) were found significantly associated in count model part. Similarly, after applying the ZINB model, no variable was found significant in zero model while sexual orientation (p-value < 0.0007), financial situation (p-value = 0.002) and happiness (p-value = 0.02) were found significant in count model part. When we took count {1, 5, 9} as inflated counts due to the presence of higher frequency in comparison with the NB and Poisson distribution (Figure 4), financial situation was not found significant in both model components (Table 5) whereas, sexual orientation (p-value < 0.015) and happiness (p-value = 0.04) were found significant only in cumulative logit model part. Inflation in counts {1, 5, 9} may have certain mechanism which needs proper investigation of population. Here these counts are takes to bring forth the possible effect of not incorporating multiple inflated counts in analysis. Moreover advice of expert of social science is needed to decide on inflated counts.

The AIC (Akaike 1974) values and BIC-corrected Vuong's (1989) test were used to compare the performance of the model. The MIGP not only has minimum AIC but also found superior based on the Vuong's test (see Tables 3 and 4).

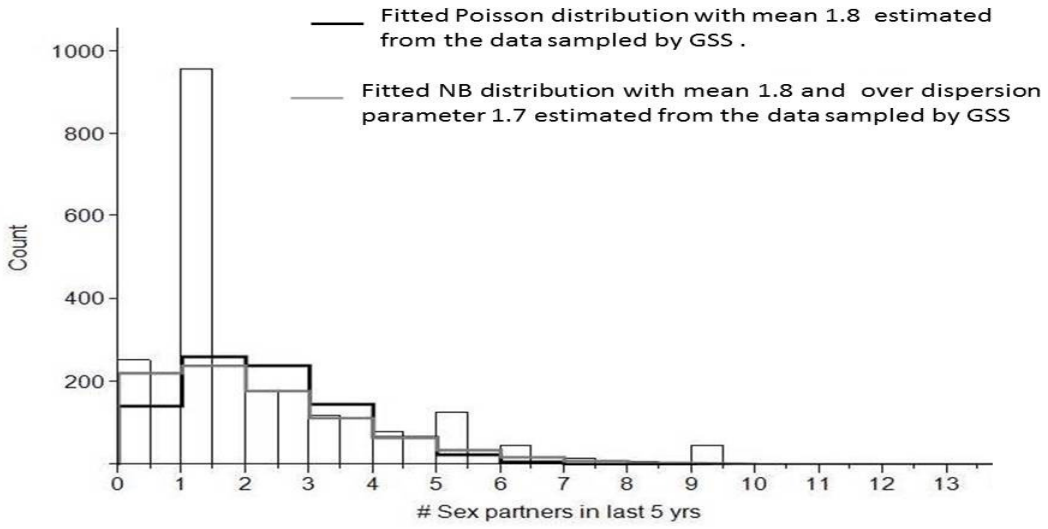


Figure 4. Histogram plot for the "number of sex partners R had in last five years" and its comparison with the Poisson distribution and negative binomial distribution with parameters estimated from the data

Table 3: The results of the Vuong's test.

Comparisons	Results	
NB vs Poisson	Test statistics (BIC-corrected)	|v|= 7.40
	P-val	P<0.0001
	Result	Favors NB
ZINB vs ZIP	Test statistic (BIC-corrected)	|v|= 7.28
	P-val	<0.0001
	Result	Favors ZINB
ZINB vs NB	Test statistics (BIC-corrected)	|v|= 9.6
	P-val	<0.0001
	Result	Favors ZINB
ZINB vs MIGP	Test statistics (BIC-corrected)	|v|= 2.37
	P-val	0.018
	Result	Favors MIGP

Table 4: AIC values.

	Poisson	NB	ZIP	ZINB	MIGP
AIC	6693.1	6339.2	6690.1	6343.3	**5515.209**

Table 5. Parameter estimates and their p-values along with the 95% CI

Parameter	Estimates	SE	z	p-value	LCI	UCI
γ_{01}	-0.632	1.160	-0.545	0.586	-2.905	1.641
γ_{02}	0.788	1.154	0.683	0.495	-1.474	3.049
$\gamma_1{}^1$	0.906	0.372	2.437	**0.015***	0.177	1.635
$\gamma_2{}^2$	0.099	0.149	0.663	0.508	-0.194	0.391
$\gamma_3{}^3$	-0.370	0.180	-2.054	**0.040***	-0.723	-0.017
α_0	0.917	0.804	1.140	0.254	-0.659	2.493
$\alpha_1{}^1$	-0.529	0.254	-2.087	**0.037***	-1.026	-0.032
$\alpha_2{}^2$	0.064	0.084	0.763	0.445	-0.101	0.229
$\alpha_3{}^3$	0.350	0.111	3.142	**0.002***	0.132	0.569
β_0	0.753	0.299	2.518	**0.012***	0.167	1.338
$\beta_1{}^1$	-0.073	0.092	-0.792	0.428	-0.252	0.107
$\beta_2{}^2$	-0.073	0.042	-1.727	0.084	-0.156	0.010
$\beta_3{}^3$	-0.076	0.053	-1.427	0.153	-0.181	0.028
ν	1.320	0.044	30.272	**<0.0001***	1.235	1.406

* Significant at the 5% level of significance, sexual orientation[1], financial situation[2], general happiness[3]

9. DISCUSSION AND CONCLUSION

We proposed a MIGP model and applied it on the simulated sets of the data as well as on the real data. We used the EM algorithm to obtain the maximum likelihood

estimates. The fact about the convergence of the EM algorithm to the local optima was also taken into the consideration so the numerical optimization was used to aid finding the global optimum. The performance of the MIGP model was evaluated and compared with the other related count models whose R packages are available. The NB and the ZINB models were primarily used for the comparisons as they always been preferred to model the over dispersed counts. The ASL was used to evaluate the performance of the MIGP and other two models. We found that the MIGP outperforms the other models.

We used a GSS data to illustrate that without including multiple inflated counts, results can be misleading and our results from the GSS data also have some implications in social science. We explored how the number of sex partners in last 5 year is associated with the financial situation and happiness of an individual. We also explored the association between sexual orientation and number of sex partners. Using our novel model, we found that financial situation is not associated with the number of sex partners in last 5 years however this predictor was significant in the ZI models and in their non-inflated analogues.

REFERENCES

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723.

Burton, A., Altman, D. G., Royston, P. and Holder, R. L. (2006) The design of simulation studies in medical statistics. *Statist. medicine,* **25**, 4279–4292.

Centers for Disease Control and Prevention (CDC) [2005 – 2006, 2007-2008, 2009-2010, 2011-2012] National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. [http://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Questionnaire]. (First Published: November, 2013)

Consul, P. C. (1989) *Generalized Poisson distributions: properties and applications*. New York: Marcel Dekker, Inc.

Consul, P. C. and Jain, G. C. (1973) A generalization of the Poisson distribution. *Technometrics*, **15**, 791-799.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B,* **39**, 1-38.

Do, C.B. and Batzoglou, S.(2008) What is the expectation maximization algorithm? *Nat Biotechnol*, **26**, 897-899.

Greene, W. H. (1994) Some Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models. Working Paper EC-94-10: Department of Economics, New York University.

GSS (2012) GSS 2010 merged with all cases and variables (Release 2, April 2012). available from [http://www3.norc.org/GSS+Website/Download/SPSS+Format/]. Accessed 11 March 2015.

Hilbe, J. M. (2011) *Negative binomial regression*. Cambridge (UK): Cambridge University Press.

JMP®, (1989-2007) Version Pro 10.0. SAS Institute Inc., Cary, NC.

Johnson, N. L., Kotz, S. and Kemp, A.W. (1992) *Univariate Discrete Distributions*. John Wiley &Sons: New York.

Lambert, D. (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34**, 1-14.

Miller, K. and Ryan, J. M. (2011) Design, Development and Testing of the NHIS Sexual Identity Question Questionnaire Design Research Laboratory, Office of Research and Methodology, National Center for Health Statistics, [http://wwwn.cdc.gov/QBANK/report/Miller_NCHS_2011_NHIS%20Sexual%20Identity.pdf]. Accessed: 30 May 2013.

NHANES (2011-2012), Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, [2001-2002] [http://www.cdc.gov/nchs/nhanes.htm] Accessed: 31 March 2014.

NHANES (2011-2012), Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Questionnaire. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, [2001-2002] [http://www.cdc.gov/nchs/nhanes.htm] Accessed: 31 March 2014.

R Development Core Team (2015) R: A Language and Environment for Statistical Computing.R Foundation for Statistical Computing, Vienna, Austria.URL http://www.R- project.org.

Schwarz, G. E. (1978) Estimating the dimension of a model. Annals of Statistics **6**, 461–464.

Smith, T. W., Marsden, P., Hout, M., and Kim, J., General Social Surveys, 1972-2012 [machine-readable data file] /Principal Investigator, Tom W. Smith; Co-Principal Investigator, Peter V. Marsden; Co-Principal Investigator, Michael Hout; Sponsored by National Science Foundation. --NORC ed.-- Chicago: National Opinion Research Center [producer]; Storrs, CT: The Roper Center for Public Opinion Research, University of Connecticut [distributor], 2013.

Su, X. G., Fan, J., Levine, R., Tan, X. and Tripathi, A. (2013) Multiple-Inflation Poisson Model with ℓ1 Regularization. *Statistica Sinica* **23**, 1071-1090.

Tallis, G. M. and Chesson, P. (1982) Identifability of mixtures. *J. Austral. Math. Soc. Ser. A* **32**, 339-348.

Vuong, Q. H. (1989) Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57, 307–333.

SUPPLEMENTARY MATERIAL

S1 Proof of Proposition 4.1

Taking $\mathfrak{F} = \{\mathbf{F}_c(\mathbf{x}); c \in \mathcal{I}\} \cup \{\mathbf{F}_c(\mathbf{x}); c \notin \mathcal{I}\}$ and noticing that the number of parameters of the MIGP depend on c. Taking

$$\mathcal{R}_1^{2\kappa} = \{\tau'_{1c}, \tau'_{2c}, \ldots \tau'_{\kappa c}; c \in \mathcal{I}\} \cup \{\tau_{1c}, \tau_{2c}, \ldots \tau_{\kappa c}; c \notin \mathcal{I}\}$$

Writing

$$H(x) = \omega_0 \cdot GP(\mu_i, \nu) + \omega_1 \cdot \mathbf{I}\{Y = c_1\} + \ldots + \omega_M \cdot \mathbf{I}\{Y = c_M\} = 0, \text{ where,}$$
$$\sum_{i=1}^{M} \omega_i = 1 \forall Y \notin I^-$$

According to Tallis et al. (1982), the above mixture $\mathcal{H}(\mathbf{x})$ is identifiable if and only if $\mathfrak{F}$ is linearly independent. We are providing the proof of this proposition by contradiction.

In order to prove this proposition, we will also use the definition of linear independence of the set of functions $\mathfrak{F}$. Tallis et al. (1982) mentioned the definition of the linear independence as follows.

A set of functions $\mathfrak{F}$ is said to be linearly independent if for real constant a_i

$\sum_{i=0}^{M} a_i F_i(x) \equiv 0 \Rightarrow a_i = 0$, for $i = 0,1,\cdots M$. More precisely,

$$a_0 \cdot GP(\mu_i, \nu) + a_1 \cdot \mathbf{I}\{Y = c_1\} + \ldots + a_M \cdot \mathbf{I}\{Y = c_M\} \equiv 0 \Rightarrow a_i = 0 \forall i \in \{0,1,\ldots,M\} \ldots 4.1$$

Suppose that $\mathfrak{F}$ is not linearly independent. Therefore,

$\sum_{i=0}^{M} a_i F_i(x) \equiv 0 \Rightarrow \exists$ at least one i such that $a_i \neq 0$.

Case-I.

Suppose that $a_i \neq 0$ for $i = 0$. Then by eq. (4.1),

$a_0 \cdot GP(\mu_i, \nu) \equiv 0 \Rightarrow GP(\mu_i, \nu) \equiv 0 \Rightarrow$ Contradiction (as per the model definition of the MIGP).

Therefore, a_0=0, this implies that $\mathfrak{F}$ is linearly independent. Therefore, by Tallis et al. (1982) $\mathcal{H}(\mathbf{x})$ is identifiable.

Case-II

Without loss of generality, suppose that $a_i \neq 0$ for any $i = 1, \cdots, M$. Then by eq. (4.1),

$a_i \cdot \mathbf{1}\{Y = c_i\} = 0 \Rightarrow \mathbf{1}\{Y = c_i\} = 0 \Rightarrow c_i \notin \mathcal{I} \Rightarrow$Contradiction (as per the model definition of the MIGP).

Therefore, $a_i = 0, \forall\, i \in \{1, 2, \cdots, M\}$ this implies that $\mathfrak{F}$ is linearly independent. Therefore, by Tallis et al. (1982) $\mathcal{H}(\mathbf{x})$ is identifiable. Therefore, the MIGP is identifiable.

A.2. Variance

$$Var(y_i) = Var\left(E(y_i|\boldsymbol{\delta}_i)\right) + E\left(Var(y_i|\boldsymbol{\delta}_i)\right)$$

$$Var\left(E(y_i|\boldsymbol{\delta}_i)\right) = \phi_i^2 \nu^2 \exp\left(\boldsymbol{b}_i^T \boldsymbol{\beta}\right)$$

$$Var(y_i|\boldsymbol{\delta}_i) = E\left(y_i^2|\boldsymbol{\delta}_i\right) - \left\{E(y_i|\boldsymbol{\delta}_i)\right\}^2$$

$$Var(y_i|\boldsymbol{\delta}_i) = E\left(y_i^2|\boldsymbol{\delta}_i\right) - \left\{E(y_i|\boldsymbol{\delta}_i)\right\}^2$$

$$= (1-\phi_i)\sum_{m=1}^{M} c_m^2 p_{im} + \phi_i\left(y_i^2|\boldsymbol{\delta}_i\{y_i \in I\}\right) - \left\{(1-\phi_i)\sum_{m=1}^{M} c_m p_{im} + \phi_i\left(y_i|\boldsymbol{\delta}_i\{y_i \notin I\}\right)\right\}^2$$

$$= (1-\phi_i)\sum_{m=1}^{M} c_m^2 p_{im} + \phi_i \left(y_i^2 | \boldsymbol{\delta}_i \{ y_i \in I \} \right) - (1-\phi_i)^2 \left(\sum_{m=1}^{M} c_m p_{im} \right)^2 - \phi_i^2 \left(y_i^2 | \boldsymbol{\delta}_i \{ y_i \notin I \} \right)$$

$$= (1-\phi_i)\sum_{m=1}^{M} c_m^2 p_{im} - (1-\phi_i)^2 \left(\sum_{m=1}^{M} c_m p_{im} \right)^2 + \phi_i (1-\phi_i) \left(y_i^2 | \boldsymbol{\delta}_i \{ y_i \notin I \} \right)$$

$$E\left(Var\left(y_i | \boldsymbol{\delta}_i \right) \right) = (1-\phi_i)\sum_{m=1}^{M} c_m^2 p_{im} - (1-\phi_i)^2 \left(\sum_{m=1}^{M} c_m p_{im} \right)^2 + \phi_i (1-\phi_i) E\left(y_i^2 | \boldsymbol{\delta}_i \{ y_i \notin I \} \right)$$

$$= (1-\phi_i)\sum_{m=1}^{M} c_m^2 p_{im} - (1-\phi_i)^2 \left(\sum_{m=1}^{M} c_m p_{im} \right)^2 + \phi_i (1-\phi_i) \left(Var\left(y_i | \boldsymbol{\delta}_i \{ y_i \notin I \} \right) + \left(E\left(y_i | \boldsymbol{\delta}_i \{ y_i \notin I \} \right) \right)^2 \right)$$

$$= (1-\phi_i)\sum_{m=1}^{M} c_m^2 p_{im} - (1-\phi_i)^2 \left(\sum_{m=1}^{M} c_m p_{im} \right)^2 + \phi_i (1-\phi_i) (\nu^2 \exp\left(\boldsymbol{b}_i^T \boldsymbol{\beta} \right)_i + \exp\left(2\boldsymbol{b}_i^T \boldsymbol{\beta} \right))$$

$$Var\left(y_i \right) = (1-\phi_i)\sum_{m=1}^{M} c_m^2 p_{im} - (1-\phi_i)^2 \left(\sum_{m=1}^{M} c_m p_{im} \right)^2 + \phi_i \exp\left(\boldsymbol{b}_i^T \boldsymbol{\beta} \right) \left(\nu^2 + \exp\left(\boldsymbol{b}_i^T \boldsymbol{\beta} \right) - \phi_i \exp\left(\boldsymbol{b}_i^T \boldsymbol{\beta} \right) \right)$$

$$Var\left(y_i \right) = E\left(y_i \right) + (1-\phi_i) \left[\sum_{m=1}^{M} c_m^2 p_{im} - (1-\phi_i) \left(\sum_{M=1}^{M} c_m p_{im} \right)^2 - \sum_{M=1}^{M} c_m p_{im} \right]$$

$$+\phi_i \exp\left(\boldsymbol{b}_i^T \boldsymbol{\beta} \right) \left(\nu^2 + \exp\left(\boldsymbol{b}_i^T \boldsymbol{\beta} \right) - \phi_i \exp\left(\boldsymbol{b}_i^T \boldsymbol{\beta} \right) - 1 \right)$$

NO ASSOCIATION BETWEEN DENTAL CARIES AND SYSTEMIC SCLEROSIS SUBTYPES

by

ARVIND TRIPATHI, HON YUEN, KUI ZHANG AND XIAOGANG SU

In preparation for Rheumatology

Format adapted for dissertation

1. ABSTRACT

Objective

The aim of this study was to determine the association between dental carries and two main subtypes of systemic sclerosis (SSc), i.e., limited and diffuse cutaneous SSc among adults.

Methods

Two novel models, namely multiple-inflation negative binomial (MINB) and multiple-inflation generalized Poisson (MIGP) models were used for modeling the counts of dental carries that typically have inflations in multiple counts, and the results were compared with existing competitive models, i.e., negative binomial (NB) and zero-inflated negative binomial (ZINB) models. Analysis to identify possible confounders was also performed and the association was explored after adjusting for the confounder Age and other covariate Income.

Results

Seventy two diffused cutaneous SSc patients and one hundred eighteen limited cutaneous patients were analyzed. In the present sample of the population of adult SSc patients, after controlling for the covariates, the zero-inflated negative binomial (ZINB) suggests that the dental caries are significantly associated with the subtypes of SSc but no significant association was found after taking into account the inflation in other counts i.e. 7 and 28.

Conclusion

This is the first study to incorporate inflation in multiple counts into the modeling of dental caries, and we found that the analysis could be misleading without taking into account the inflations in the counts other than zero. Therefore, we strongly recommend incorporating the inflation present in any counts in the analysis along with zeroes.

2. INTRODUCTION

Systemic sclerosis (SSc, scleroderma), one of a group of an autoimmune connective tissue diseases, is characterized by vascular dysfunction, inflammation and excessive fibrosis of connective tissues supporting the skin and visceral organs [1, 2]. SSc is classified into two main subtypes: limited cutaneous SSc (lcSSc) is characterized by restricted skin involvement and slow progression; diffuse cutaneous SSc (dcSSc) is characterized by rapid progression, visceral organ and symmetrical widespread skin involvement [3].

SSc affects the skin and musculoskeletal tissues along with the oral and perioral tissues [4-7]. A major clinical manifestation, orofacial dysfunction (e.g., microstomia and xerostomia) may lead to oral health problems among people with SSc [8-10]. Excessive dry mouth resulting from salivary hypofusion is known as Xerostomia which not only promotes the development of dental plaque but also increases the risk of the development of the dental caries [11-13]. Recently, Yuen et al. performed a cross-sectional study to identify factors associated with increased gingival inflammation in adults with SSc [10]. The increased gingival inflammation represents the poor oral health and the authors found SSc subtype a significant factor of gingival inflammation. Oral clinical conditions also affect our daily performance [14].

Baron et al. found that SSc patients have more dental caries and periodontal disease when compared with the general population [15]. In another study consisting of 163 SSc patients (72% with limited and 28% with diffuse cutaneous disease), Baron et al. found that tooth loss is associated with poor upper extremity function, gastro-esophageal reflux disease (GERD) and decreased saliva [16]. In the first study, the authors compared

the SSc patients with the general population, but the second study consisted of only SSc patients. However, the association between the dental caries and SSc subtypes remained unaddressed. Mahjour et al. investigated and discussed the association of SSc with its possible oral manifestations and concluded that early diagnosis of oral symptoms of SSc is extremely important [17].

The DMFT index is commonly used in dental research and is regarded as the "gold standard" to measure the cumulative dental caries. **DMFT** index represents the total number of Decayed (**D**), Missing due to caries (**M**), and Filled (**F**) permanent **T**eeth. Thus, count data arise from DMFT index are gathered and analyzed for oral health problem. In epidemiological studies, Coxe et al. characterized a count as the number of occurrences of an event during a certain period of time [18]. The maximum score of DMFT could be either 32 or 28 (depending on the inclusion of the third molars); the minimum score is 0.

Statistical modeling plays an important role in understanding dental caries, their development and the associated risk factors. In determining the appropriate model choice, several aspects need to be considered. First of all, the presence of over dispersion characterized by the greater population variance than the mean should be acknowledged. The subsequent discussion is provided only to bring forth a few studies reflecting the dispersion present in the dental caries data and the application of the NB to model them [19-21]. For example, Grainger and Reid in 1954 worked on the distribution of the dental caries in children and found that the NB satisfactorily describes the distribution [19]. They also suggested that "this is in agreement with the idea that the dental caries is a chance phenomenon with individuals differing in their susceptibility to tooth decay" [19].

Thitasomakul et al. in 2009 also applied the NB distribution to model risk for early childhood caries [20]. Brennan et al. in 2007 also used the NB distribution to model the dental caries for indigenous adult public dental patients in Australia [21].

Diesendorf mentioned that with the mandated addition of fluoride to the drinking water in many communities, significant reductions in the prevalence of dental caries in the last few decades have been observed [22]. As a result, the presence of many zero counts have been observed in the dental caries data. Therefore, zero inflated models were applied extensively for analysis; for example, Javali et al. applied zero inflated count models to model the dental caries [23]. Böhning et al. applied the zero-inflated Poisson (ZIP) model to analyze DMFT counts [24], whereas Mwalili et al. applied the zero inflated negative binomial (ZINB) model with correction for misclassification in the dental caries research [25].

Despite the popular use of the zero inflated models, inflation in other DMFT counts is frequently observed, for example, the high prevalence of dental caries among older adults (age 65+) was shown by an average score of 18 in the DMFT index (United States, National Health and Nutrition Examination Survey, 1999–2004) [26]. A high count of 28 is also prevalent due to the fact that fitting a full denture requires taking out all teeth. The above facts indicate that in the population under study, not only 0 but also other counts such as 28 could also have inflation. No study provides the analysis to address the presence of such inflation in the counts other than the zero counts mainly due to unavailability of the appropriate statistical models to address such issues. The commonly used count models such as the Poisson and negative binomial (NB) do not provide adjustment for the inflated counts. While the NB model can be used for over

dispersed (i.e. Population variance is greater than the mean) counts, the equidispersion assumption (i.e. population variance is equal to mean) of Poisson model make it more restrictive in use. A generalized Poisson model can be applied in over/ under dispersed (i.e. population variance is less than mean) counts and mitigate the restrictive assumption of equidispersion of Poisson model, but this solution does not provide adjustment for inflated counts. Similarly, zero inflated analogues of the Poisson, NB and generalized Poisson models, i.e. ZIP, ZINB and ZIGP provide the adjustment for excess in zero counts but remain silent when other than zero count is inflated. Fortunately, some recent developments in the count modeling open the new avenues to model the count data with multiple inflated counts. Recently, Su et al. (2013) has proposed multiple-inflation Poisson (MIP) model to address the issues of multiple inflated counts present in equidispersed counts[27], and two novel models the multiple-inflation negative binomial (MINB) and the multiple-inflation generalized Poisson (MIGP) have also been proposed [28] to handle situations when multiple inflated counts are present with highly dispersed non-inflated counts. The superiority of the MINB and MIGP models over ZINB model was also found to analyze the multiple inflated counts in a simulated data set [28]. Considering the presence of the multiple inflated counts along with highly dispersed non inflated DMFT counts, we applied MINB and MIGP model to explore any true significant association between DMFT counts and SSc subtypes.

As indicated in the literature, adults with SSc were at greater risk for oral disease, and dental caries are more prevalent among SSc population than the general population. However, no studies have investigated the disease severity as indicated by the type of SSc and the dental caries. Knowing whether the severity of SSc disease (as indicated by

its type) has any impact on dental caries may help clinicians better prepare themselves when educating their patients and researchers better understand the complex manifestations of disease severity on oral disease in this population. Therefore, one specific aim of the present analysis is to determine whether or not there is a significant association between dental caries and the subtypes of SSc. The association of the dental caries with many other severe health conditions and/ or socioeconomic consequences cannot be denied. Therefore covariate adjustment is critical in terms of teasing out possible confounding effects on association of SSc subtypes with dental caries.

2. METHOD

2.1. Participants

The study included adults more than eighteen years of age who were diagnosed with SSc and who fulfilled the preliminary classification criteria of American College of Rheumatology for SSc [29]. However, the individuals with the localized scleroderma (e.g. morphea, linear scleroderma, and en coup de sabre) were excluded and not considered eligible for the participation.

2.2. Recruitment

Study participants were recruited through the Medical University of South Carolina (MUSC) scleroderma clinic and a local connective tissue disease database (CTDD). The CTDD, a database of medical information of SSc patients, contains information about the majority of patients who received treatment and/or consultation from the MUSC scleroderma clinic beginning in 2001. In the start of the study (October, 2007), the contact information of 509 SSc patients was obtained from the CTDD; This information was used to invite them to participate in the dental survey study. To

participate in the study, the patients were contacted via phone or personally by the research coordinator for participation on behalf of the physicians at the MUSC scleroderma clinic. The study was explained by the research coordinator to the potential participants. Information about their time commitment, obtaining verbal consent and scheduling an oral examination appointment was provided. A single dental visit of about 1.5 hours duration was required in the study.

2.3. Procedure

After obtaining the informed consent, the oral examination was conducted by two trained and calibrated dental hygienists at the MUSC General Clinical Research Center. The examination included measurement of the oral aperture, assessment of manual dexterity to perform oral hygiene, and dental and periodontal health, as well as completion of the Center of Epidemiological Studies Depression (CES-D) Scale (30), a self-report instrument to assess depressive symptomatology, and a package containing an oral health-related questionnaire. A mouth cavity assessment (31) was also conducted. The protocol was approved by the MUSC Institutional Review Board. The clinical trial protocol number was NCT01817361.

2.4. Statistical Methods

We applied the NB, ZINB, MINB, MIGP models for modeling the data. The NB model has been applied when the outcome variable (e.g. DMFT) was over dispersed. The over dispersion means population mean is greater than population variance. When there is equality in population mean and variance (referred as equidespersion) then it reduces into the Poisson model by itself. However, the ZINB has been applied when outcome variable was over dispersed and also presented an excess of the count zero. Here excess

means more than expected counts or inflated counts. Moreover, the MINB can be applied when the outcome variable is over dispersed with multiple inflated counts. On the other hand, the MIGP could also be applied when outcome variable is under dispersed with multiple inflated counts. In the NB, we model mean by using loglinear model. However, in ZINB along with loglinear model for mean of the regular counts, excess of zeros are modeled using logit model. Moreover, in the MINB and MIGP, inflated counts are modeled using cumulative logit model along with loglinear model for the mean of the regular counts and logit model is also used to model the mixing probability of mixing the inflated counts with the regular counts. For variable selection, L1 regularization method was used to select the variables in the NB model and step wise variable selection method was used to select the variables in logit model.

3. DATA ANALYSIS

To assess the association of DMFT with two main subtypes of SSc, a data set of 190 patients with SSc was used. The list of variables considered for the analysis is provided in Table 1. Among the 190 observations, only 169 (72 diffused cutaneous SSc patients and 118 limited cutaneous patients) had complete cases, i.e. without missing values, which were considered for the analysis. The histogram plot for the distribution of the DMFT along with its comparison with the NB distribution is given in Figure 1, whereas Figure 2 provides a comparison of the DMFT with the Poisson distribution. JMP Pro 10 statistical software (SAS Institute Inc., Cary, NC) [32] was used to create histogram plots of DMFT. The discrete fit option of the JMP was used to compare it with the NB (Gamma Poisson) and the Poisson distributions. The R programming language [33] was used in the development of the MINB and MIGP models. The application of the

NB model, ZINB model, Vuong's test, minimum AIC criterion and the generation of Figures 3-6 were also performed using R.

Table 1: Description of variables used in the analysis

No.	Var Name	Description
1	Diag	1 if diffuse cutaneous SSc and 2 if limited cutaneous SSc
2	dental_visit	1 if within the past year (1 to 12 months ago), 2 if within the past 2 years (1 to 2 years ago), 3 if within the past 5 years (2 to 5 years ago), 4 if 5 or more years ago and 5 if don't know/not sure/never
3	floss_evening	1 if yesand 2 if no
4	Drymouth	1 if yes and 0 if no
5	M_aperture	Mean oral apartue
6	Siccasymptoms	1 if yes and 0 if no
7	Sjogren	1 if yes and 0 if no
8	Age	Age in years
9	Race_minority	1 if non -white and 2 if white / caucasian
10	GenderNm	1 if male and 2 if female
11	marital_status	1 if married, 2 if living with partner / not legally married, 3 if divorced / separated, 4 if widowed and 5 if single / never married
12	Education	1 if grade school, 2 if high school, 3 if technical school, 4 if college(BA/BS), 5 if some or complete graduate school
13	Employment	1 if currently full time job, 2 if currently part time job, 3 if unemployed, 4 if retired, 5 if homemaker and 6 if disability
14	income	1 if the household's total combined income in the last year before taxes is below $10,000, 2 if the household's total combined income in the last year before taxes is between $10,000 –$ 14,999, 3 if the

		household's total combined income in the last year before taxes is between $15,000 – $19,999, 4 if the household's total combined income in the last year before taxes is between $20,000 – $24,999, 5 if the household's total combined income in the last year before taxes is between $25,000 – $29,999, 6 if the household's total combined income in the last year before taxes is between $30,000 – $34,999, 7 if the household's total combined income in the last year before taxes is between $35,000 – $39,999, 8 if the household's total combined income in the last year before taxes is between $40,000 – $44,999, 9 if the household's total combined income in the last year before taxes is between $45,000 –$ 49,999, 10 if the household's total combined income in the last year before taxes is between $50,000 – $54,999, 11 if the household's total combined income in the last year before taxes is between $55,000 – $59,999, 12 if the household's total combined income in the last year before taxes is between $60,000 – $64,999, 13 if the household's total combined income in the last year before taxes is between $65,000 – $69,999, 14 if the household's total combined income in the last year before taxes is between $70,000 – $74,999 and 15 if the household's total combined income in the last year before taxes is $75,000 and over
15	Smoking	1if regularly smoker, 2 if occasional smoker, 3 if former smoker and 4 if non smoker
16	chew_gum	In the past week, number of days of chewing gum.
17	sugar_snack_days	In the past week, number of days of consuming sugar-containing snacks.
18	sugar_snack_freq	Frequency of eating sugar-containing snacks past days.
19	sweet_drink_days	In the past week, how many days did you drink regular non-diet soda/sweet tea between meals
20	soda_freq	Number of cans of 12 oz regular non-diet soda consumed last day
21	sweet_tea_freq	Number of glasses of regular sweet tea drunk last day

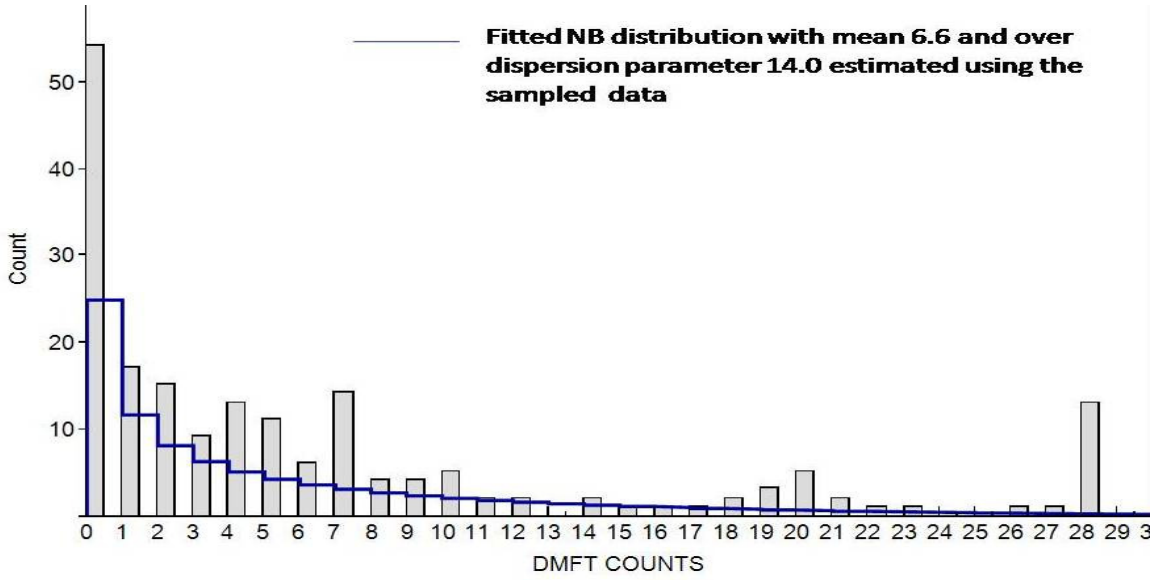


Figure 1. Histogram plot for the DMFT counts and its comparison with the negative binomial distribution.

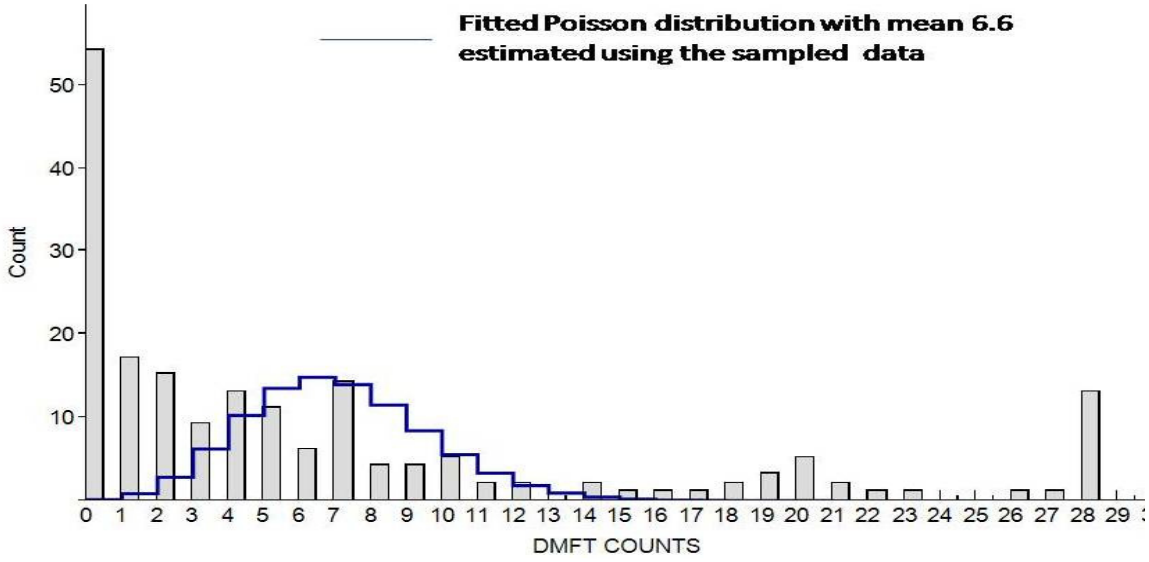


Figure 2. Histogram plot for the DMFT counts and its comparison with the Poisson distribution.

Location and the over dispersion parameter estimates (along with the 95% CI) of the NB distribution obtained after using the discrete fit option were 6.59 (5.38, 8.17) and 14.0 (10.64, 19.10) respectively. However, the estimate of the scale parameter of the

Poisson distribution (along with the 95% CI) was 6.59 (6.23, 6.96). These two histogram plots (Figures 1 and 2) suggest that the NB distribution fits better to DMFT counts than the Poisson distribution. Moreover, when we compared the fitted NB distribution with the distribution of DMFT, the inflation in the counts 0, 7 and 28 was easily observed at the histogram plot (Figure 1). Therefore, $I = \{0, 7, 28\}$ was considered as a set of counts with inflation. Since no formal test for the inflation in multiple counts is available, it depends on researcher's discretion to decide on the inflation in the counts and it is also research area specific. The presence of even fewer inflated counts may need to be adjusted based on the research field and the researcher hypothesis. However sample specific variation should always be taken into account and only those counts should be considered inflated which are present in the relatively higher frequencies. Moreover, the proper investigation of the data prior to analysis is always recommended, and for this purpose histogram plot has obvious advantages and is used frequently. Therefore, we recommend use of the histogram plot to find the inflation in the counts.

To find the possible confounders, the association of the variables provided in the list in Table 1 with DMFT as well as with SSc subtypes was explored. The variables associated with the both, i.e., with DMFT and with SSc subtypes were considered as confounders. The summary statistics of the categorical and ordinal variables (given in Table 1) across different type of SSc is provided in Tables 2 and 3 respectively. However, the mean (standard deviation) of the continuous variables Age and mean oral aperture was 55.07 years (13.03 years) and 37.9 mm (8.4 mm) respectively. The analysis is performed after adjusting for the confounders.

Table 2: The summary statistics for categorical (i.e. nominal) variables across subtypes of SSc

Name of Variable		Diffuse cutaneous SSc n (%)	Limited cutaneous SSc n (%)
Dental visit	Within the past year (1 to 12 months ago)	55 (76.39)	86 (73.50)
	Within the past 2 years (1 to 2 years ago)	5 (6.94)	12 (10.26)
	Within the past 5 years (2 to 5 years ago)	4 (5.56)	10 (8.55)
	5 or more years ago	6 (8.33)	7 (5.98)
	if don't know/not sure/never	2 (2.78)	2 (1.71)
Floss evening	Yes	33 (46.48)	51 (44.74)
	No	38 (53.52)	63 (55.26)
Dry mouth	Yes	19 (27.14)	31 (27.68)
	No	51 (72.86)	81 (72.32)
Sjögren's syndrome	Yes	3 (6.4%)	7 (9.7%)
	No	44 (93.6%)	65 (90.2%)
Sicca Symptoms	Yes	8 (12.3%)	31 (31.3%)
	No	57 (87.7%)	68 (68.7%)
Race	Non white	34 (47.22)	31 (26.27)
	White / Caucasian	38 (52.78)	87 (73.73)
Gender	Male	22 (30.56)	9 (7.63)
	Female	50 (69.44)	109 (92.37)
Marital Status	Married	43 (59.72)	68 (58.12)

	Living with partner / not legally married	2 (2.78)	5 (4.27)
	Divorced / separated	11 (15.28)	17 (14.53)
	Widowed	2 (2.78)	17 (14.53)
	Single / Never Married	14 (19.44)	10 (8.55)
Education	Grade School	3 (4.17)	3 (2.56)
	High school	23 (31.94)	30 (25.64)
	Technical school	20 (27.78)	17 (14.53)
	College (BA/BS)	21 (29.17)	43 (36.75)
	Some or complete graduate school	5 (6.94)	24 (20.51)
Employment	Currently full time job	16 (22.22)	33 (28.21)
	Currently part time job	6 (8.33)	9 (7.69)
	Unemployed	3 (4.1)	2 (1.71)
	Retired	11 (15.28)	42 (35.90)
	Homemaker	6 (8.33)	6 (5.13)
	Disability	30 (41.67)	25 (21.37)
Smoking	Regularly smoker	2 (2.78)	4 (3.42)
	Occasional smoker	3 (4.17)	5 (4.27)
	Former smoker	13 (18.06)	30 (25.64)
	Non smoker	54 (75.00)	78 (66.67)

Table 3: The summary statistics for categorical (i.e. ordinal) variables across subtypes of SSc.

Name of variables	Diffuse cutaneous SSc Median ($Q1^*$,$Q3^*$)	Limited cutaneous SSc Median ($Q1^*$, $Q3^*$)
Income	7 (3, 13)	10 (4, 15)
In the past week, number of days of chewing gum.	0 (0, 3)	0 (0, 3)
In the past week, number of days of consuming sugar-containing snacks.	4 (2, 7)	4 (2, 7)
Frequency of eating sugar-containing snacks past days.	1 (1, 2)	1 (0, 2)
In the past week, how many days did you drink regular non-diet soda/sweet tea between meals.	2 (0, 6)	1 (0, 3)
Number of cans of 12 oz regular non-diet soda consumed last day	0 (0, 1)	0 (0, 1)
Number of glasses of regular sweet tea drunk last day	0 (0, 1.5)	0 (0, 1)

*Q1 = Lower quartile, Q3= Upper quartile

The TCOUNTREG (count regression, SAS 9.3 (SAS Institute))[34], penalizing the likelihood with L1 norm, procedure was used for the variable selection in the NB model to model DMFT counts. Penalizing the likelihood provides the variable selection by shrinking the coefficient to zero. The selected variables were then used in the NB model part of the ZINB and MINB model. The TCOUNTREG procedure is useful in analyzing regression models in which the dependent variable takes nonnegative integer or count values (SAS Institute). The variables given in Table 1 are used for the variable selection. Since the presence of collinearity and missingness across the variables hinders the variable selection process by stopping the optimization, only important demographic, disease related and eating habit related factors are considered.

The variables Age $\left(t\ (df=1)\ =\ 5.69,\ P<0.0001\right)$ and Income $\left(t\ (df=1)\ =\ -5.24,\ P<0.0001\right)$ were selected with the optimal tuning parameter 0.45. The dispersion parameter was estimated to be 0.85 and found to be significantly $\left(t\left(df=1\right)=\ 5.51,\ P<0.0001\right)$ different from zero. For all the tests, the statistical significance was set at $\alpha=0.05$. The presence of dispersion was in accordance with the finding of the histogram plot (Figure 1) where the counts were over dispersed and the NB distribution was the best fit. Due to the presence of over dispersion, only NB, ZINB, MINB and MIGP models were used to model DMFT counts in order to assess its true association with the SSc subtypes.

As mentioned earlier, the confounding variable is a covariate that is associated with both the dependent (i.e. DMFT in the present study) and the independent (i.e. SSc subtype in the present study) variables. We found that only Age and Income are associated with the DMFT, so either of them that has association with the SSc subtype would be a confounder. For this purpose, the logistic regression model was used by taking SSc subtypes as the dependent variable and Age and Income as the independent variables. Only the variable Age was significantly ($Wald\,\chi_1^2=8.43, P=0.0037$) associated with the SSc subtypes. As a result, among the variables mentioned in Table 1, only Age was found to be associated with both the DMFT and SSc subtype and hence considered a confounder.

The commonly known fact that has also been mentioned by Arlene Flink (2008), "Confounder can lead to the false conclusion that the dependent variable is in the casual relationship with the independent or the predictor variables" [35]. Since Age was a confounder, the true association between DMFT and SSc subtypes was explored after

adjusting for the variable Age. Since in the variable selection for the NB model discussed earlier, other than the Age, variable Income was also found significantly associated with response variable DMFT, so Income was also included as covariate in the MI-models.

Further analysis was performed by taking DMFT as a response variable, SSc subtype as a predictor, Age as a confounder and Income as a covariate. For the ZINB, MINB and MIGP models, DMFT as dependent variable and SSc subtypes, Age and Income as independent variables were taken to model both the inflated counts and regular DMFT counts following NB/ GP distribution.

In order to decide on the best model among the NB, ZINB, MINB and MIGP models, the minimum Akaike information criterion, i.e., AIC (Akaike 1974) [36] along with Vuong's (1989) non nested test [37] were used, due to the presence of the non-nested models. However, due to the presence of the three model components, the multiple- inflation models involve the estimation of several more parameters than the zero inflated models and their non-zero inflated analogs. This creates a bias in Vuong's test in favor of the more complex model. To address the issue, we used the Schwarz correction[38].

After using Vuong's test and comparing the NB with the ZINB model, the ZINB model was preferred over the NB model. Further, the ZINB model was compared with the MINB and MIGP models. In the ZINB model, DMFT was found to be significantly associated with the SSc subtypes in the count model part (P=0.0175) as well as in the zero model part (P=0.0185) (See Table 4 for the parameter estimates and their P-values). However, after applying the multiple inflation models, i.e. MINB and MIGP, the variable SSc subtypes was found no longer significantly associated with the DMFT (See Table 4

for the parameter estimates and their P-values). This suggests that incorporating other than zero counts as inflated can affect the significance of the result. If inflation in the other counts is not incorporated, the results could be inaccurate and could thus mislead researchers.

Table 4: Parameter estimates and their p-values.

Model		Variables	Parameter estimates	SE	P-Value
ZINB	Count Model	Intercept	-0.03	0.45	0.94
		Age	0.04	0.01	<0.0001*
		Income	-0.10	0.02	<0.0001*
		diag	0.37	0.16	**0.02***
	Zero Model	Intercept	-0.86	1.53	0.57
		Age	-0.06	0.02	0.002*
		Income	0.05	0.05	0.31
		diag	1.45	0.61	**0.02***
MINB	Count Model	Intercept	0.74	0.43	0.09
		Age	0.03	0.01	<0.001*
		Income	-0.11	0.02	<0.0001*
		diag	0.09	0.15	**0.55**
		nu	2.29	0.71	<0.0001*
	Model for counts with inflation. (Cumulative logit Model)	Intercept 1	5.76	2.17	0.01*
		Intercept 2	6.89	2.28	<0.0001*
		Age	-0.12	0.04	<0.0001*
		Income	0.19	0.08	0.01*
		diag	0.29	0.82	0.72
	Model for Mixing Probability. (logit Model)	Intercept	-0.87	0.90	**0.33**
		Sweet drink days	0.24	0.09	0.01*
		Employment	0.06	0.10	0.53
		Age	0.02	0.02	0.27
MIGP	Count Model	Intercept	1.20	0.44	0.01*
		Age	0.03	0.01	<0.0001*
		Income	-0.10	0.02	<0.0001*
		diag	0.07	0.16	0.68
		nu	2.35	0.34	<0.0001*
	Model for counts with inflation. (Cumulative logit Model)	Intercept 1	5.55	2.23	0.01*
		Intercept 2	7.10	2.38	<0.0001*
		Age	-0.11	0.04	<0.0001*
		Income	0.15	0.08	0.05*
		diag	0.29	0.85	0.73
	Model for Mixing Probability. (logit Model)	Intercept	-1.30	0.97	**0.18**
		Sweet drink days	0.25	0.10	0.01*
		Employment	0.07	0.10	0.47
		Age	0.02	0.02	**0.15**

*significant at the o.o5 level of the significance

The mixing probability in the multiple inflation count models was used to obtain the mixture of the discrete distribution and the NB/generalized Poisson distribution. The logistic regression was used to obtain the mixing probability. For this purpose, a dichotomous variable with the value 0 for the non-inflated DMFT counts and 1 for the inflated DMFT counts was used as the response variable. The explanatory variables in the logistic regression model were selected by using the stepwise variable selection method. For this purpose, SAS 9.3's (SAS Institute) logistic procedure with the option of stepwise variable selection was used. The variables Age, Employment and "number of days of drinking sweet drinks in the past week" are entered in the model (with the probability to enter and probability to stay at 0.05) and used as the explanatory variables to model the mixing probability.

4. RESULTS AND DISCUSSION

As indicated in the histogram plot (Figure 1), the inflation in the counts 7 and 28 along with the inflation in the count 0 were obvious. In Figures 5-6, the MINB and the MIGP models provide better fit to the data when compared to the ZINB and the NB models. Figures 3-6 provide the histogram plot of the DMFT counts superimposed on the fitted value of the best-fit NB, ZINB, MINB and MIGP models. Vuong's test and the minimum AIC criterion were used to select the best model. The AIC of the NB, ZINB, MIGP and MINB model were 999.05, 980.48, 862.76 and 861.01 respectively. Therefore, based on minimum AIC criterion MINB was found the best model. However, only a slight difference in the AIC across MINB and MIGP model was found.

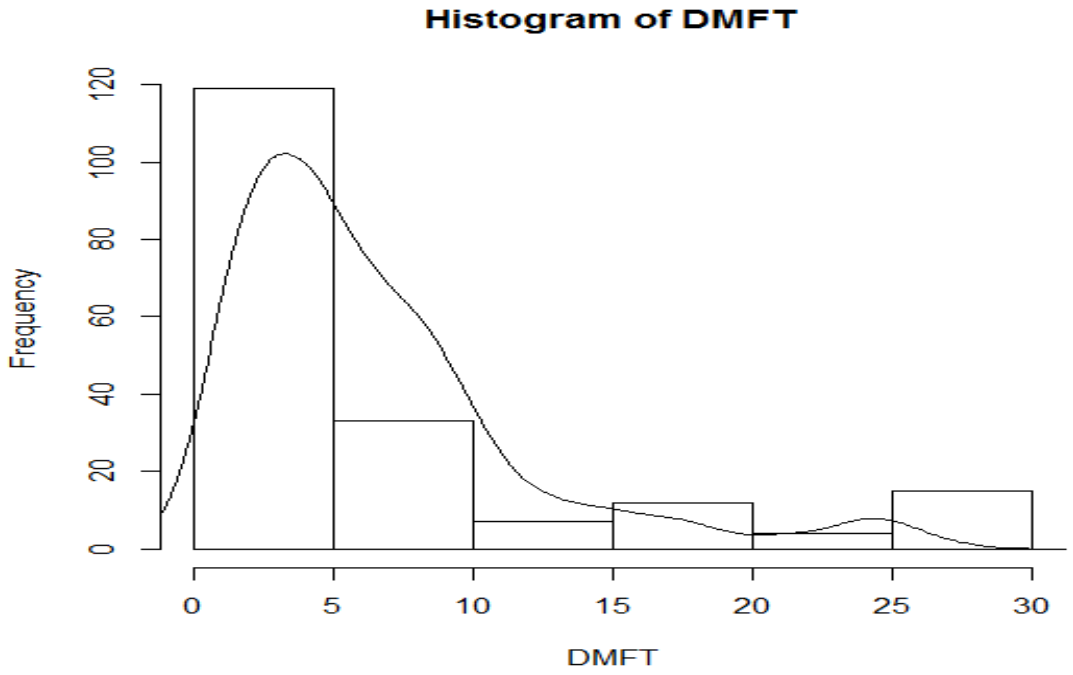


Figure 3. Histogram plot of the DMFT counts superimposed with the fitted values obtained from the NB model.

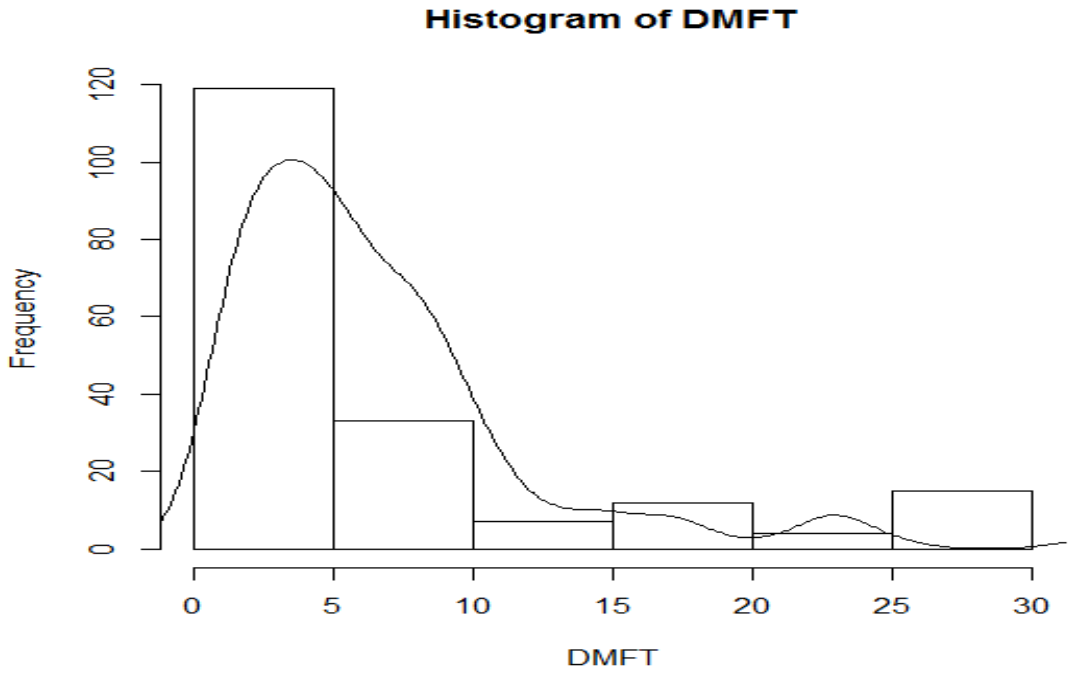


Figure 4. Histogram plot of the DMFT counts superimposed with the fitted values obtained from the ZINB model.

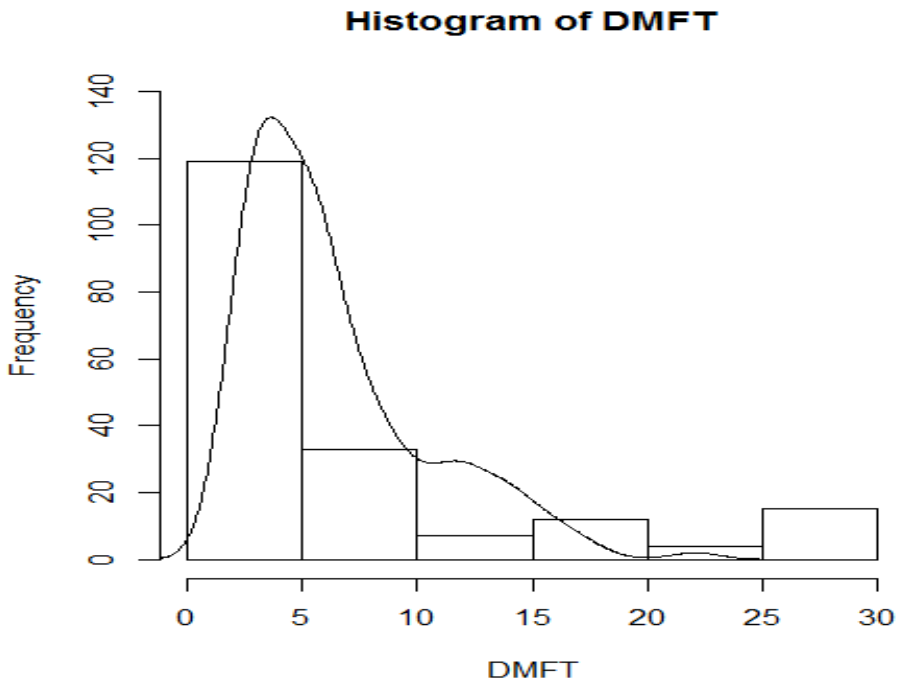


Figure 5. Histogram plot of the DMFT counts superimposed with the fitted values obtained from the MINB model.

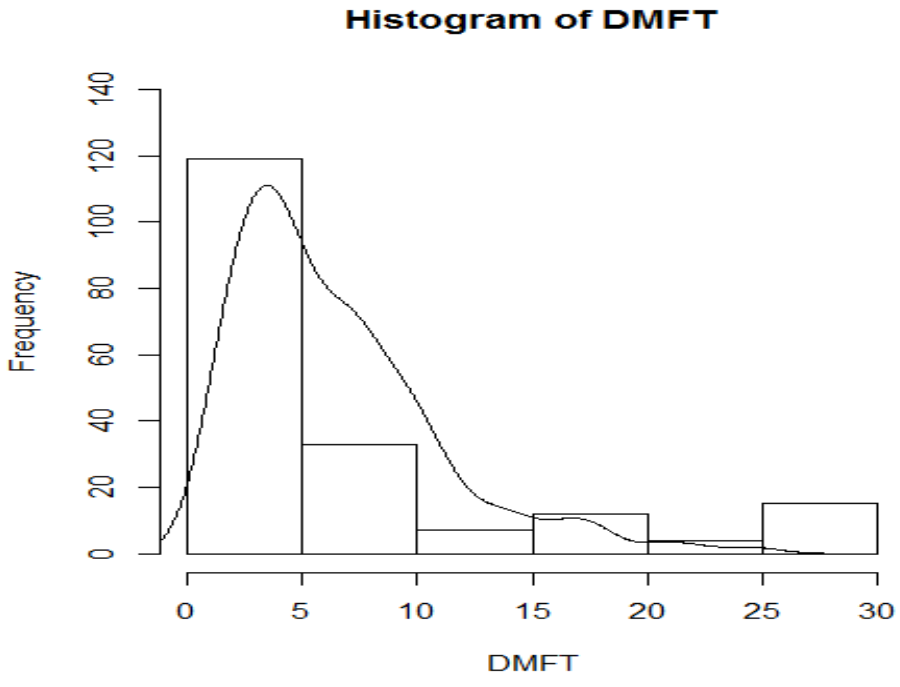


Figure 6. Histogram plot of the DMFT counts superimposed with the fitted values obtained from the MIGP model.

Similar results were obtained by using Vuong's test. Vuong's test significantly (Test statistics $|\nu| = 2.51,\ P < 0.006$) favors the ZINB in comparison to the NB model. However, both the MIGP (Test statistics $|\nu| = 11,\ P < 0.0001$) and the MINB models (Test statistics $|\nu| = 11,\ P < 0.0001$) were preferred over ZINB. Moreover, neither MIGP nor MINB were significantly (Test statistics $|\nu| = 0.667,\ P < 0.50$) preferable over each other.

In the present sample of the population of the adults with SSc, we found that SSc subtypes are not associated with DMFT. We also found that we can easily get misleading results if we consider only the zero count as inflated. This was in accordance with the fact that not considering the significant information provided by the data and applying only the existing methods for sake of the convenience of the analysis could easily provide misleading results.

Although no formal test for identifying the inflation in counts is available, a thoughtful inspection of the data before analysis (e.g. inspecting the histogram plot) is always recommended and could reveal the inflation in counts if any.

5. STUDY LIMITATIONS

The small size of sample is one of the major limitations of the study. Due to the presence of three model parts in the multiple inflation models, the larger the sample size, the more precise the results.

REFERENCES

1. Derk C T, Jimenez S A. Systemic sclerosis: current views of its pathogenesis. Autoimmun Rev 2003; 2: 181-91.

2. Dghoughi S, El Wady W, Taleb B. Systemic sclerosis. Case report and review of literature. N Y State. Dent J 2010; 76: 30-5.

3. Chung L, Lin J, Furst D E, Fiorentino D. Systemic and localized scleroderma. Clin Dermatol 2006; 24: 374–92.

4. Alantar A, Cabane J, Hachulla E, et al. Recommendations for the care of oral involvement in patients with systemic sclerosis. Arthritis Care Res (Hoboken) 2011; 63: 1126-33.

5. Albilia J B, Lam D K, Blanas N, et al. Small mouths ... Big problems? A review of scleroderma and its oral health implications. J Can Dent Assoc 2007; 73: 831-36.

6. Fischer D J, Patton L L. Scleroderma: oral manifestations and treatment challenges. Spec Care Dentist 2000; 20: 240-4.

7. Skare T L, Toebe B L, Boros C. Hand dysfunction in scleroderma patients. Sao Paulo Med J 2011;129: 357-60.

8. Poole J L, Brewer C, Rossie K, et al. Factors related to oral hygiene in persons with scleroderma. Int J Dent Hyg 2005; 3: 13-7.

9. Scardina G A, Messina P. Systemic sclerosis: description and diagnostic role of the oral phenomena. Gen Dent 2004; 52: 42-7.

10. Yuen H, Weng Y, Reed S, et al. Factors associated with gingival inflammation among adults with systemic sclerosis. Int J Dent Hyg 2014; 12: 55-61.

11. Guggenheimer J, Moore P A. Xerostomia: etiology, recognition and treatment. J Am Dent Assoc 2003; 134: 61-69.

12. Hase J C, Birkhed D. Salivary glucose clearance, dry mouth and pH changes in dental plaque in man. Arch Oral Biol 1988; 33 :875-80.

13. Sreebny L M, Valdini A. Xerostomia. A neglected symptom. Arch Intern Med 1987;147: 1333-7.

14. Gomes A S, Abegg C, Fachel J M. Relationship between oral clinical conditions and daily performances. Braz Oral Res. 2009; 23: 76-81.

15. Baron M, Hudson M, Tatibouet S, et al. The Canadian systemic sclerosis oral health study: orofacial manifestations and oral health-related quality of life in systemic sclerosis compared with the general population. Rheumatology 2014; 53: 1386-94.

16. Baron M, Hudson M, Tatibouet S, et al. The Canadian systemic sclerosis oral Health Study III: Relationship between disease characteristics and oro-facial manifestations in systemic sclerosis. Arthritis Care Res (Hoboken) 2014 published on 2014 Oct 9. doi: 10.1002/acr.22490.4.

17. Mahjour F, Hosseini F S, Talebi A M R, et al.The Relationship between Systemic Sclerosis And Periodontal Disease. 2012; 7th IADR Iranian Division Annual Meeting.

18. Coxe S, West S G, Aiken L S. The analysis of count data: A gentle introduction to Poisson regression and its alternatives. J Pers Assess 2009; 9: 121-36.

19. Grainger R M, Reid D B W. Distribution of dental caries in children. J Dent Res 1954; 33: 613–23.

20. Thitasomakul S, Piwat S, Thearmontree A, et al. Risks for early childhood caries analyzed by negative binomial models. J Dent Res 2009; 88:137-41.

21. Brennan D S, Roberts-Thomson K F, Spencer A J. Oral health of indigenous adult public dental patients in Australia. Aust Dent J 2007; 52: 322-28.

22. Diesendorf M. The Mystery of Declining Caries. Community Dent Oral 1989; 17 : 106-7.

23. Javali S B, Pandit P V. Using zero-inflated models to analyze dental caries with many zeroes, Indian J Dent Res 2010; 21: 480-5.

24. Böhning D, Dietz E, Schlattmann P, et al. The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. J Roy Statist Soc Ser. A 1999; 162: 195–09.

25. Mwalili S M, Lesaffre E, Declerck D. The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research. Stat Methods in Med Res 2008; 17:123-39.

26. National Institute of Dental and Craniofacial Research: Dental Caries (Tooth Decay) in Seniors (Age 65 and Over). United States, National Health and Nutrition Examination Survey, 1999–2004.Available at: http://www.nidcr.nih.gov/DataStatistics/FindDataByTopic/DentalCaries/DentalCariesSeniors65older.htm. Accessed: February 4, 2015.

27. Su X G, Fan J, Levine R, et al. Multiple-Inflation Poisson Model with L1 Regularization. Statistica Sinica 2013; 23: 1071-90.

28. Tripathi A. (2015). Count Models with Multiple Inflations (Unpublished doctoral dissertation). University of Alabama at Birmingham, Birmingham, Alabama, US.

29. Masi A T. Preliminary criteria for the classification of systemic sclerosis (scleroderma). Subcommittee for scleroderma criteria of the American Rheumatism Association Diagnostic and Therapeutic Criteria Committee. Arthritis Rheum 1980; 23 : 581-90.

30. Radloff L S. The CES-D Scale: a self-report depression scale for research in the general population. Appl Psychol Meas 1977; 1: 385-401.

31. Hill E G, Slate E H, Wiegand R E, et al. Study design for calibration of clinical examiners measuring periodontal parameters. J Periodontol 2006; 77: 1129-41.

32. JMP®, Version Pro 10.0. SAS Institute Inc., Cary, NC, 1989-2007.

33. R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

34. SAS Institute Inc. 2011. Base SAS® 9.3 Procedures Guide. Cary, NC: SAS Institute Inc.

35. Fink A. Practicing Research: Discovering Evidence That Matters. Sages Publication Inc, 2008; 100.

36. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control 1974; 19: 716–23.

37. Vuong Q H. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 1989; 57:307–33.

38. Schwarz G E. Estimating the dimension of a model. Annals of Statistics 1978; 6: 461–4.

CONCLUSION

Summary

In medical research, it is important to choose appropriate analytical tools to model the outcome/response variable and predictors in order to find the significant predictors associated with the progression or remission of diseases. The choice of appropriate models depends on the distribution of outcome/response variables. When the response variable is a count, the Poisson distribution or the negative binomial (NB) distribution is generally used to characterize the response variable. The Poisson distribution is appropriate when outcome variable is sampled from a population having the variance same as the mean (i.e. an equidispersed population), whereas the NB is appropriate when the population variance is more than the mean (i.e. an over dispersed population). The equidispersion property of the Poisson distribution makes it very restrictive to use, and thus the amendment in the distribution to generalize it was made earlier by many researchers. As a result different generalizations of the Poisson distribution are proposed. The generalized Poisson distribution is found useful when the outcome variable is sampled from a population having the population variance either less or more than the mean (over/under dispersed counts). Observing some counts in higher frequencies than expected under the above discussed distributions is common in real world situations. These counts are referred as inflated counts. The presence of inflated counts itself is responsible for inducing over / under dispersion in the data. The mixture of distributions

is used to model such a count outcome appropriately and precisely. When the frequency of the zero count was higher than expected, the zero-inflated models have been extensively used. However, when the inflation appears in multiple counts other than zero, we realized the lack of appropriate models for the analysis. This problem is the focus of the present research. Specifically, in the current research, the multiple inflated (MI) models such as multiple-inflation negative binomial (MINB) and multiple-inflation generalized Poisson (MIGP) models are proposed for over/ under dispersed response variable with multiple inflated counts. Moreover, the one step smoothly clipped absolute deviation (SCAD) is also adapted to aid in the variable selection in the MI models.

In particular, in the first paper, we developed the MINB model and assessed its performance with simulated data. To obtain the maximum likelihood estimates, we used the expectation maximization (EM) algorithm along with the numerical optimization. A three-step procedure is adapted which involves the combination of the application of the truncated NB model, cumulative logit model, logit model, numerical optimization and the EM algorithm. The performance of the MINB model was compared with the other competitive count models which are frequently used to model the over dispersed data such as the NB and ZINB models along with the log-linear and ZIP models. The average square loss (ASL) is used to evaluate the performance of the MINB and the other models. We found that the MINB performs better than the other above mentioned models in simulated data. The presence of more than one model component in the zero and multiple-inflation count models makes the variable selection also an important issue which is also addressed in the first paper. The most widely used variable selection procedures—the forward, backward, bidirectional sequential testing method and the best

subset selection method—become computationally prohibitive when the variables are large. The SCAD penalty does not only aid in variable selection but it also offers better estimates; hence, it is used in this paper along with the local linear approximation (LLA) of the penalty. In particular, we used one step SCAD variable selection method and found that it has high sensitivity and moderate specificity. We defined specificity and sensitivity as correctly identifying the zero covariates as zero and correctly identifying non-zero coefficient as non-zero respectively.

In the second paper, we developed the MIGP model and evaluated its performance with simulated data. To obtain the maximum likelihood estimates, we again used the EM algorithm along with the numerical optimization. The performance of the MIGP model was compared with the other related count models. The NB and ZINB models were mainly used as the candidates for the comparisons as they always were preferred to model the over dispersed counts. The ASL was used to evaluate the performance of the MIGP and other two models. We found that the MIGP outperforms the other models.

In the third paper, the newly developed models (MINB and MIGP) were applied to a real data set consisting of 190 systemic sclerosis (SSc) patients. The data contain seventy two diffused cutaneous SSc patients and one hundred eighteen limited cutaneous patients. The DMFT (i.e. decayed, missing and filled teeth) is a count variable and one of the most commonly used indicators of the oral health and was thus used as the response variable in our analysis. The objective of this paper was to find any significant associations between the DMFT and SSc subtypes. In addition to the SSc subtypes, age and income were used as covariates, since both of them were found associated with the

DMFT. In the observed data, the DMFT is found highly dispersed along with multiple inflated counts. The zero-inflated negative binomial (ZINB) was used and the results suggested that the DMFT was significantly (p= 0.02) associated with the subtypes of the SSc at the 5% level of significance, but no significant association was found after taking into account the inflation in other counts, i.e. 7 and 28. We find that there is no significant association between SSc subtypes and the DMFT. We also find that we can easily get misleading results if the inflation at multiple counts is not appropriately modeled.

Future Research

This research could be expanded in the future in several ways.

Development of a test for multiple inflations

The presence of the multiple inflated counts can easily be recognized by using histogram plot, but the development of a formal test would be of great value to settle on the inflation in the counts. In addition, a test to detect the over dispersion could also be extended for the presence of multiple inflated counts in future.

Mixed model

In statistical analysis, we often encounter correlated data, predominantly due to the grouping of the subjects, due to the repeated measures on each subject over the time or space, or due to the multiple related outcome measures at one point in time. The presence of such correlation makes the multiple-inflation count modeling even more challenging. However, neither ignoring the grouping entirely nor fitting each group with separate model is recommended. The mixed effect model could be used to analyze such

data. Therefore one of our future projects is to develop a mixed effect model that can be used for correlated multiple-inflated count data.

Bayesian models

Incorporating the prior information about parameters can sometimes lead to a better estimation and inference. Since no Bayesian analogue to handle such an issue of multiple inflations is available, Bayesian multiple inflated count models could be considered for future research.

Multiple inflated count models for spatial data

The presence of multiple inflated counts is often observed in environmental data that are spatial and temporal. The inflation of the zero counts in such a data has been found and zero inflated spatial models have been proposed. Therefore, the development of the multiple inflation count spatial models will be our future extension.

Multiple inflated truncated count models

In our research, we encountered many examples in which multiple counts were present but the logical structure of the variable excluded certain counts, typically zeros. This suggests the need to develop multiple inflated truncated count models.

GENERAL LIST OF REFERENCES

Akaike, H. (1974). A new look at model identification. *IEEE Transactions an Automatic Control* **19**:716-723.

Buu, A., Johnson, J. N. J., Li, R. and Tan, X. (2011). New Variable Selection Methods for Zero-Inflated Count Data with Applications to the Substance Abuse Field. *Statistics in Medicine* **30**:2326-2340.

Cameron, A. C. and Trivedi, P. K. (1986). Econometric models based on count data: Comparisons and applications of some estimators. *Journal of Applied Econometrics* **1**: 29-53.

Chen, L. S., Hutter, C. M., Potter, J. D., Liu, Y., Prentice, R. L., Peters, U. and Hsu L. (2010). Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. *The American Journal of Human Genetics* **86**:860-871.

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*, *Series B* **39**: 1-38.

Dodge, Y. (2008). The Concise Encyclopedia of Statistics. Springer Science and Business Media, LLC.

Eggenberger, F. and Polya, G. (1923). Uber die Statistik verketteter Vorgange. *Zeitschrift für Angewandte Mathematik und Mechanik* **1**, 279-289.

Efroymson, M. A. (1960). Multiple regression analysis. Mathematical Methods for Digital Computers. Wiley, New York.

Farewell, V. T. (1986). Mixture models in survival analysis: Are they worth the risk? *Canadian Journal of Statistics* **14**: 257-262.

Feuerverger, A. (1979). On Some Methods of Analysis for Whether Experiments. *Biometrika* **66**: 655-658.

Frank, I. E., and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics* **35**:109-148.

Furnival, G. and Wilson R. (1974). Regression by leaps and bounds. *Technometrics* **16**:499-511.

Gardner, A. (2011). Pack-a-day smokers declining. Health.com March 15, 2011 4:44 p.m. EDT.

Gosset, S.W. ("Student") (1907). On the error of counting with a haemacytometer. *Biometrika* **5**:351-360.

Greenwood, M. and Yule, G. U. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal Statistical Society, Series A* **83**:255-279.

Greene, W. H. (1994). Some Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models. Working Paper EC-94-10: Department of Economics, New York University.

Hall, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: A case study. *Biometrics* **56**: 1030-1039.

Hall, D. B. and Shen, J. (2010). Robust estimation for zero-inflated Poisson regression. *Scandinavian Journal of Statistics* **37**: 237-252.

Heilbron, D. C. (1994). Zero-altered and other regression models for count data with added zeros. *Biometrical Journal* **36**:531-547.

Hilbe, J. M. (2011). Negative binomial regression. Cambridge (UK): Cambridge University Press.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **12**: 55-67.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics* **35**: 73-101.

Jochmann, M. (2009).What Belongs Where? Variable Selection for Zero-Inflated Count Models with an Application to the Demand for Health Care. The Rimini Centre for Economic Analysis, WP 09-45.

Ladislaus, J. B. (1898). Das Gesetz der kleinen Zahlen [The law of small numbers]. Leipzig, Germany: B.G. Teubner.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34**: 1-14.

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.

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.

Meeker, W. Q. (1987). Limited Failure Population Life Tests: Application to Integrated Circuit Reliability. *Technometrics* **29**: 151-165.

Montmort, P. R. de (1713). Essai d'analyse sur les jeux de hasard. Quillau, Paris 2nd edn.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics* **33**: 341-365.

Nelder, J. and Wedderburn, R. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society, Series A* **135**: 370-384.

NHANES (2001-2002), Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, [2001-2002] [http://www.cdc.gov/nchs/nhanes.htm] Accessed: 11 August 2013.

NHANES (2001-2002), Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Questionnaire. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, [2001-2002] [http://www.cdc.gov/nchs/nhanes.htm] Accessed: 11 August 2013.

Ni, X. S. and Huo, X. (2006). Regressions by enhanced leaps-and-bounds via optimality tests (LBOT). Technical report, Georgia Inst. Technology. Available at www2.isye.gatech.edu/statistics/papers/. Accessed: 11 August 2013.

Pascal, B. (1679). Varia Opera Mathematica. D.Petri de Fermat. Tolosae.

Pearson, K. (1894). Contributions To The Mathematical Theory of Evolution. *Philosophical Transactions of the Royal Society* **185**: 71-110.

Poisson, S. D. (1837). Probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilitiés, Paris.

Preisser, J.S., Stamm, J.W. and Long, D. L. (2012). Review and Recommendations for Zero- Inflated Count Regression Modeling of Dental Caries Indices in Epidemiological Studies. *Caries Research* **46**:413-423.

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http: //www.R-project.org/.

Genuer, R., Morlais, I., and Toussile,W. (2011). Gametocytes infectiousness to mosquitoes: variable selection using random forests, and zero inflated models. Research Report RR-7497, INRIA, URL http://hal.inria.fr/inria-00550980/en/.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**: 461-464.

Staub, K. E. and Winkelmann, R. (2012). Consistent estimation of zero-inflated count models. *Health Economics* **22**: 673–686.

Su, X. G., Fan, J., Levine, R., Tan, X., and Tripathi, A. (2013). Multiple-Inflation Poisson Model with L1 Regularization. *Statistica Sinica* **23**: 1071-1090.

Su, X. G. (2014). Variable Selection via Subtle Uprooting. *Journal of Computational and Graphical Statistics.* (Accepted).

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics - Theory and Methods* **A7**: 13-26.

Tallis, G. M. and Chesson, P. (1982). Identifability of mixtures. *Journal of the Australian Mathematical Society, Series A* **32**:339-348.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B* **58**: 267-288.

Teicher, H. (1961). Identifiability of mixtures. *The Annals of Mathematical Statistics* **32**: 244-248.

Yau, K. K. W., Wang, K. and Lee, A. H. (2003). Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with Extra Zeros. *Biometrical Journal* **45**:437-452.

Yuen, H. K., Weng, Y., Bandyopadhyay, D., Reed, S. G., Leite, R. S, and Silver, R. M. (2011).Effect of a Multi-Faceted Intervention on Gingival Health Among Adults with Systemic Sclerosis. *Clinical and Experimental Rheumatology* **29**(2 Suppl 65): S26-S32.

APPENDIX A

INSTITUTIONAL REVIEW BOARD APPROVAL

THE UNIVERSITY OF ALABAMA AT BIRMINGHAM

Institutional Review Board for Human Use

DATE: August 9, 2013

MEMORANDUM

TO: Arvind Tripathi
Principal Investigator

FROM: Cari Oliver, CIP
Assistant Director
Office of the Institutional Review Board (OIRB)

RE: Request for Determination—Human Subjects Research
IRB Protocol #N130802004 – Multiple Inflated Negative Binomial Models

A member of the Office of the IRB has reviewed your Application for Not Human Subjects Research Designation for above referenced proposal.

The reviewer has determined that this proposal is **not** subject to FDA regulations and is **not** Human Subjects Research. Note that any changes to the project should be resubmitted to the Office of the IRB for determination.

470 Administration Building
701 20th Street South
205.934.3789
Fax 205.934.1301
irb@uab.edu

The University of
Alabama at Birmingham
Mailing Address:
AB 470
1720 2ND AVE S
BIRMINGHAM AL 35294-0104