1-1-2013

Heaped Data In Count Models

Tammy Harris
University of South Carolina

Heaped Data in Count Models

by

Tammy Reneé Harris

Bachelor of Science
Presbyterian College, 2005

Master of Science in Public Health
University of South Carolina, 2007

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Biostatistics

The Norman J. Arnold School of Public Health

University of South Carolina

2013

Accepted by:

James W. Hardin, Major Professor

James R. Hussey, Committee Member

Alexander C. McLain, Committee Member

Kevin J. Bennett, Committee Member

Lacy Ford, Vice Provost and Dean of Graduate Studies

Dedication

This dissertation is dedicated to my parents who have always encouraged me to do my best in everything I do and instilled that 'go-get-it' attitude in me. Thank you for believing in me and giving me your support throughout the years.

Acknowledgments

I would like to express my sincere gratitude to my major advisor, Dr. James Hardin, for challenging me, editing many drafts, reviewing code and results with me until they were correctly stated, and constant encouragement. Thank you for being patient and not kicking me out of your office when I had many questions to ask, you have given me insight into several fields to deepen my understanding of count models. To the remainder of my committee members, Drs. James Hussey, Alexander McLain, and Kevin Bennett, I express my sincere thanks for your patience and assistance in the dissertation process.

To my parents, Tommy and Kathy Harris, thank your for always sharing an encouraging word and patience throughout this process. You have always supported me and told me to simply to trust in God and everything will work out in His time. To my brother, Michael, thanks for the pep talks and never doubting my abilities and knowledge. To my fiancé Christian, thanks for the constant support and encouragement as I completed this work. You have all given me unlimited support during my graduate study, no matter how stressful or challenging it was, and I thank you. I love you all.

Abstract

Heaped data result when subjects who recall the frequency of events prefer for reporting from a limited set of rounded responses or preferred digits over reporting exact counts. These rounded responses and digit preferences (also referred to as data coarsening) could be characterized by reported frequencies (or counts) favoring multiples of 20, reporting counts ending with 0 or 5, or a preference for reporting an even number over an odd number or vice versa. This mixture of values is a type of measurement error (pattern of misreporting) that can lead to biased estimation and imprecision in discrete quantitative data. Sometimes this pattern in data can be explained or understood, but its effect on the statistical inference may be harder to anticipate. A visual representation of heaped data can be seen in a frequency distribution (histogram) where the heaps are represented as periodic peaks or spikes within the overall data layout. Some common examples of heaped count data include smoking (cigarette) cessation studies, blood pressure (BP) measurements, unemployment duration data, reported age, reported weight, frequency of sexual intercourse, breastfeeding months, number of required menstrual cycles before pregnancy, and reported birth weight.

We develop statistical models to model heaped count data using a mixture of likelihood functions for heaped and nonheaped count data. For the heaped count data, we consider that the reported outcome is actually censored over the half width of the heaping multiple. Simultaneously, we consider that nonheaped data follow the count distribution's likelihood for exact counts; that they are not censored. The investigator specifies the heaping multiples over which heaped values are censored via an interval regression approach in our approach. We also create new Stata commands

to model these heaped data and use real world data as well as simulated data to illustrate our approach.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In many medical applications, reported count data (frequencies of events, symptoms, behaviors, etc.) are rounded to reflect preferential selection from a limited set of numbers. Preference for reporting from a limited response set is referred to as heaping. Heaped data occur when subjects eschew reporting exact counts in favor of reporting counts from a limited response set, multiples of common vales, according to a preferred digit. Such rounded responses and digit preference (also referred to as data coarsening) could include multiples of 5 or 10, or selecting an even number over an odd number. When some data reflect an exact count, and other data are heaped, the mixture of values represents reporting error (pattern of misreporting) that can lead to biased estimation and imprecision in discrete quantitative data. Sometimes heaping patterns in data can be explained or anticipated, but its effect on the statistical inference may be more difficult. Heaped data can be seen in a frequency distribution (histogram) where heaps appear as periodic peaks or regularly spaced spikes within the overall data layout.

1.1 Literature Review

A source of heaped count data results from cigarette cessation studies where the respondent reports the number of cigarettes smoked in a specific time period (Wang and Heitjan [2008];Klesges et al. [1995];Lewis-Esquerre et al. [2005]). Participants of these types of smoking studies, tend to round their reported cigarette counts to multiples of 5, 10, or 20 which may be due to the number of cigarettes in a quarter pack, half pack, or pack, respectively. Another example where heaping can occur is in the collection of blood pressure (BP) measurements for which there is a commonly seen terminal digit preference (Nietert et al. [2006]). In this data heaping is exemplified when BP readings tend to be recorded in measurements ending in 0 or 5 and even numbers preferred over odd numbers. Other examples of heaped data include frequency of sexual intercourse, breastfeeding months (Roberts and Brewer [2001]), duration data for unemployment (Wolff and Augustin [2003]), self-reported age (Pardeshi [2010]), number of required menstrual cycles or months before pregnancy (Ridout and Morgan [1991]), and reported birth weight (Channon et al. [2011]).

Relevant literature include different approaches to addressing heaped count data. For instance, Wang and Heitjan [2008] proposed a Bayesian proportional odds rounding behavior model that was a function of the unobserved true count value and a latent heaping behavior variable. This latent heaping behavior variable took into account four values of cigarette counts and rounded the data: exact counts, multiples of 5, multiples of 10, and multiples of 20 (size of cigarette pack). The authors then compared the model fits and performed model selection in a Bayesian approach by using certain prior distributions and Bayes factors to estimate parameters in the posterior distribution. These authors analyzed only univariate count data (no covariates) and only considered heaping at multiples of 5. Heitjan and Rubin [1990] filled-in (imputed) correct ages for data that contained 270 Tanzanian children from Dodoma. They estimated the rounding probabilities given the observed data and imputed the ages based on that and assumed that the ages of children were associated with different types of rounding. Those types of rounding behaviors occurred in exact age, age rounded to the nearest half-year, and age rounded to the nearest full-year. Their method of analyses used was multiple imputation with simple and more complex models. Thavarajah et al. [2003] encountered heaping for BP readings with preference of 0, 2, 4, 6, 8, or an odd number as the measurements last digit. The authors used a χ^2 test to examine the tendency for zero digit preference and nonzero digit preference for certain demographic information, along with a logistic regression to analyze zero bias.

Roberts and Brewer [2001] introduced two more simple, yet general approaches to heaping in discrete count data. One approach referred to as "‘neighbor difference"’ used the differences between the frequency of the response and the mean of the two neighboring (nearest) frequency responses. The other approach, called local mode, takes into account the mode of the response in binary fashion. The authors propose using the sum of the values for either choice (neighbor difference or local mode) for a set of responses. Their method allows for a measure of the magnitude of heaping and hypothesis testing for the presence of heaping in the data by a p-value. For both approaches to work, the investigative team would need to start with some hypothesized heaped (multiple) values and maximum discrete count response based on the given data. The authors demonstrate their approach on an interviewed study regarding the number of drug partners each subject had. The interviewers questioned the subjects two different ways, one method using a numerical estimate and the other using a partner elicitation method. The numerical method simply required subjects to estimate the number of drug partners, while the partner elicitation method required subjects to recall drug partners individually and then count each partner. Based on the data, heaping was apparent for multiples of 5. The authors then compared the two methods (numerical and elicitation) while using their proposed heaping analysis and concluded that by using the subjects’ estimation of drug partners, heaping was more likely to be apparent than using the partner elicitation method.

Digit preference or heaping in the study of fecundity arises from retrospective reporting of womens time-to-pregnancy (TTP), which are commonly rounded to 6, 12 or even 3 cycles (Ridout and Morgan [1991]). Ridout and Morgan [1991] assumed that the TTP data had an underlying beta-geometric distribution. Under this assumption, they showed that this heaped data did not change the conclusions from fitting a beta-geometric distribution, but absorbed the lack of fit. With similar data, Price and Seaman [2006] used a computationally intensive approach, that used Markov chain Monte Carlo (MCMC) methods, to model fecundity. These authors took a Bayesian approach by using a hierarchically centered generalized linear mixed model, but the MCMC method was complicated and limited use of the full conditional distribution. This model was able to compute posterior distributions and interval estimates of all of the regression parameters.

Some studies that used naive approaches included a study involving rural areas in India where a door-to-door open-ended questionnaire was used for data collection (Pardeshi [2010]). Digit preference and age heaping were shown in these data with a preference for ages ending in 0, 5, or both. The author used Whipple's index and Myers' blended index to measure age preference but these indices exclude childhood and old age respectively. The Whipple's Index measures the extent of preference for ages ending in 0, 5, or both by using age responses between 23 and 62 and calculating a value based on Whipple's Index formula which has a minimum value of 0 and maximum of 500. A value of 0 indicates that ending digits 0 and 5 are not reported, a value 100 indicates no preference for ending digits of 0 and 5, while a value of 500 indicates that ending digits of 0 and 5 were always reported. Pardeshi [2010] describes the index accuracy for the age distribution as: < 105 = highly accurate; $105 - 109.9$ = fairly accurate; $110 - 124.9$ = approximate; $125 - 174.9$ = rough; and ≥ 175 = very rough. While the Myers' blended index (Myers [1940]), considers preferences of ages ending in any of the ten digits (0 to 9), creating an overall age accuracy score. This index assumes that the population is equally distributed among the different ages and has a mininum value of 0 (no heaping) and a maximum value of 90 (same reported ending digit for entire population). The Whipple index was also used to measure age heaping in cancer patients (Denic et al. [2004]) and amongst women (married versus unmarried) using population surveys spanning 400 years by (Foldvari et al. [2012]). The Whipple's Index has a major limitations: only considering preferences of measures for 2 digits, 0 and 5; considers only the interval of ages between 23 and 62; handle only a single year worth of data. The Myers' Blended Index have no theoretical basis, is not suitable for group data, and does not take into account any other forms of heaped bias. Lastly, Channon et al. [2011] just calculated percentages of low birth weight (LBW) in retrospective studies from 6 developing countries. Preference of rounding (from memory or recorded health card) was found to occur in the nearest digit for birth weight in multiples of 100g and 500g. The authors state the goal of the study was not to propose a method to more accurately redistribute the birth weights heaped on 2,500g, but to demonstrate the effect of heaping on LBW estimates. These studies lacked the use of statistical modeling and predictions.

Chapter 2 describes our new method for handling heaping in count data and also introduces new interval-censored regression models that may be used to analyze heaped count data in Section 2.1. The Data Analysis chapter will exhibit 4 sets of analyses using our new method by analyzing data from an NHANES study in Section 3.1 of cigarette counts, a National Institute of Mental Health (NIMH) multisite HIV (human immunodeficiency virus)/STD (sexually transmitted disease) prevention trial of frequency of sexual activity in Section 3.2, a simulation study involving heaped poisson data in Section 3.3, and finally score test derivatives for interval-censored regression models are discussed in Section 3.4. In Chapter 4 we give a complete discussion of our results and future work.

Chapter 2

Methods

We propose statistical models to model heaped count data using a mixture of likelihood functions for heaped and nonheaped count data. We also create new heaped count data regression commands in Stata statistical software. We consider the reported outcome is actually censored over the half width of the heaping multiple for heaped count data. We also consider that nonheaped (not censored) data follow the count distribution's likelihood for exact counts. For example, for heaped data which are heaped at multiples of 20, the counts that are reported of non-multiples of 20 will be treated as exact results, such that $P(Y \in \{y - \lfloor 20/2 \rfloor, y + \lfloor 20/2 \rfloor\})$ for those counts with multiples of 20, instead of $P(Y = y)$ for exact counts. The investigator should specify the heaping multiples over which heaped values are censored via an interval regression approach for our new method.

2.1 Models

For the following models, let y_{Li}, y_{Ri} represent the right and left endpoints of interval-censored count observations, respectively. We have

$$\begin{aligned} y_{Li} &= \max\{0, y_i - \lfloor h_i/2 \rfloor\} \\ y_{Ri} &= y_i + \lfloor h_i/2 \rfloor \\ h_i &= h_j = \max_{j=1,\ldots,H} I(y_i \bmod h_j = 0) \end{aligned}$$

where $\lfloor h_i/2 \rfloor$ is the half width of the heaping interval, and $h_1 = 1$. If all observations are exact (noncensored, no heaping), then $H = 1$ and these formulas simplify to that of Poisson, generalized Poisson, and Negative Binomial regression.

Poisson Model

Poisson regression analysis is often used to analyze response variables comprising count data. This distribution describes the probability of the number of event occurrences and the expected number of occurrences modeled through explanatory variables. For a random variable Y_i, we have a response vector $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, where n is the sample size and Y_i, Y_j are independent and identically distributed (iid) for any $i \neq j$. An invertible link function is used to describe the relationship between the linear predictor $x_i\beta = \eta_i$ to the expected value of the responses μ_i via $\mu_i = \exp(\mathbf{x}_i\beta)$ where $\mathbf{x}_i$ is a covariate vector and β is a vector of regression parameters to be estimated. The probability mass function is given by

$$f(y_i;\mu_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}, y_i = 0, 1, 2, \ldots, \quad \mu_i > 0. \tag{2.1}$$

The Poisson model has some strong assumptions, one being equidispersion; that is, that the mean (μ_i) and variance (μ_i) of the outcomes are equal for a given set of covariates. When the variance exceeds the mean (overdispersion), or the variance is smaller than the mean (underdispersion), the Poisson assumption is violated.

In practice, equidisperson $Var(y)/E(y) = 1$ is rarely reflected in data and using the Poisson model which carries this assumption leads to poor estimates of the variance, and, thus, to poor inference. In most situations, the variance ($Var(y)$) exceeds the mean ($E(y)$) for a given count variable Y. This occurrence of extra-Poisson variation is known as overdispersion $Var(y)/E(y) > 1$ (see, for example, Dean [1992]). Lee and Nelder [2000] describe two approaches to model overdispersed count data

(i) Quasi-likelihood approach;

(ii) Include a random-effect model

where the quasi-likelihood approach involves the extension of the parametric model by extra parameters to allow for a more general variance structure. In situations for which the variance is smaller than the mean, data are characterized as being underdispersed. Puig and Valero [2006] state that dispersion is a measure of departure detection from the Poisson distribution which can be examined in various ways, Fisher overdispersion test, zero-inflation index, etc. Modeling overdispersed or underdispersed count data using inappropriate models can lead to underestimated or overestimated standard errors and misleading inference.

We use coefficient estimates and standard errors to obtain the maximum likelihood method. For a random sample of observations $y_1, y_2, \ldots, y_n$, we know that the Poisson log-likelihood function is

$$\mathcal{L} = \sum_{i=1}^{n} \left\{ y_i \ln(\mu_i) - \mu_i - \ln \Gamma(y_i + 1) \right\} \tag{2.2}$$

We propose, for a random sample of heaped observations, the log-likelihood function (interval-censored regression) is given by

$$\begin{aligned} p_{1i} &= \Gamma_I\{y_{Li}, \mu_i\} = 1 - P(Y \leq y_{Li} - 1 | Y \sim Poisson) \\ p_{2i} &= \Gamma_I\{y_{Ri} + 1, \mu_i\} = 1 - P(Y \leq y_{Ri} | Y \sim Poisson) \end{aligned}$$

where Γ_I is the regularized incomplete gamma function and y_{Li}, y_{Ri} represent the right and left censored observations, as stated above, respectively. Therefore, the interval-censored Poisson regression model has the log-likelihood equal to

$$\mathcal{L} = \sum_{i=1}^{n} \ln(p_{1i} - p_{2i}). \tag{2.3}$$

Generalized Poisson Model

Consider the generalized Poisson (GP) distribution having a probability mass function

$$f(y_i; \mu_i, \alpha) = \frac{\mu_i(\mu_i + \alpha y_i)^{y_i - 1} e^{-\mu_i - \alpha y_i}}{y_i!}, \quad y_i = 0, 1, 2, \ldots \tag{2.4}$$

where α is the dispersion parameter, $\mu_i > 0$, $\max(-1, \frac{-\mu_i}{4}) < \alpha < 1$, and $\mu_i = \exp(x_i\beta)$. When $\alpha \to 0$, the GP distribution reduces to the Poisson distribution. This distribution can be used to analyze equidispersed, overdispersed, or underdispersed count data. The mean and variance for the generalized Poisson distribution (Hardin and Hilbe [2012]) is

$$\begin{aligned} E(Y_i) &= \frac{\mu_i}{1-\alpha}, \text{and} \\ Var(Y_i) &= \frac{\mu_i}{(1-\alpha)^3}. \end{aligned}$$

For a random sample of observations $y_1, y_2, \ldots, y_n$, the GP log-likelihood function is

$$\mathcal{L} = \sum_{i=1}^{n} \Big\{ \ln \mu_i + (y_i - 1)\ln(\mu_i + \alpha y_i) - \mu_i - \alpha y_i - \ln \Gamma(y_i + 1) \Big\}. \tag{2.5}$$

Consul and Famoye [1992] and Consul [1989] extensively studied this distribution and illustrated that covariates can be introduced into a regression model via the relationship

$$\log \frac{\mu_i}{1-\alpha} = \sum_{r=1}^{p} x_{ir}\beta_r, \tag{2.6}$$

where x_{ir} is the ith observation of rth covariate, p is the number of covariates in the model, and β_r is the rth regression parameter. We propose, for a random sample of heaped observations, the log-likelihood function (interval-censored regression) is given by

$$\begin{aligned} p_{1i} &= \Gamma_I\{y_{Li}\alpha, \mu_i\} \\ p_{2i} &= \Gamma_I\{(y_{Ri}\alpha) + 1, \mu_i\} \end{aligned}$$

where Γ_I is the regularized incomplete gamma function. Therefore, the log-likelihood function suitable for heaped data using a GP model is

$$\mathcal{L} = \sum_{i=1}^{n} \ln(p_{1i} - p_{2i}). \tag{2.7}$$

Negative Binomial Model

Suppose we have count data that is overdispersed or underdispersed, therefore a Poisson regression model is not appropriate. Therefore, a model that's been extensively used by researchers over time, the negative binomial distribution (Lawless [1987];Dean and Lawless [1989]) is considered. In each trial th e probability of success is p and of failure is $(1-p)$. The general probability mass function of the negative binomial (NB) distribution is

$$f(y;\alpha,p) = \frac{\Gamma(y+\frac{1}{\alpha})}{\Gamma(y+1)\Gamma(\frac{1}{\alpha})} p^{1/\alpha}(1-p)^y, \quad y = 0,1,2,\ldots \tag{2.8}$$

where α is the dispersion parameter. When $\alpha \to 0$, this reduces to the Poisson distribution. The mean and variance for the negative binomial distribution is as follows

$$\begin{aligned} E(Y_i) &= \frac{1-p}{\alpha p}, \text{and} \\ Var(Y_i) &= \frac{1-p}{\alpha p^2} \\ &= \frac{p-p^2+p^2-2p+1}{\alpha p^2} \\ &= \frac{p(1-p)+(p-1)^2}{\alpha p^2}. \end{aligned}$$

The negative binomial can be altered by using the log-linear specification $g(x;\beta) = \exp(x^T\beta)$ (Lawless [1987]) where x is the p x 1 vector of explanatory variables and β is a vector of regression parameters. Lawless [1987] states that a Poisson model would stipulate that the distribution of $Y|x$ is Poisson with mean equal to $\mu(x) = T\{g(x;\beta)\}$. Based on this information, the negative binomial regression model is

$$f(y_i;\alpha,\mu_i(x)) = \frac{\Gamma(y_i+\frac{1}{\alpha})}{\Gamma(y_i+1)\Gamma(\frac{1}{\alpha})}\left(\frac{1}{1+\alpha\mu_i(x)}\right)^{1/\alpha}\left(\frac{\alpha\mu_i(x)}{1+\alpha\mu_i(x)}\right)^{y_i}, \quad y_i = 0,1,2,\ldots, \tag{2.9}$$

where α is the dispersion parameter. Here we use a common re-parameterization using $p = \frac{1}{1+\alpha\mu}$ where p relies on covariates x, which results in the mean and variance of Y as

$$\begin{aligned} E(Y|x) &= \mu(x) \\ Var(Y|x) &= \mu(x) + \alpha\mu^2(x) \end{aligned}$$

Therefore, we have $Y \sim NB(\mu(x), \alpha)$. Again, when $\alpha \to 0$ the mean and variance reduces to the the mean and variance of a Poisson model. For a random sample of observations $y_1, y_2, \ldots, y_n$, the NB log-likelihood function is

$$\mathcal{L} = \sum_{i=1}^{n} \Bigg\{ \ln(\Gamma(y_i + 1/\alpha)) - \ln(\Gamma(y_i + 1)) - \ln(\Gamma(1/\alpha)) + (1/\alpha)\ln\left(\frac{1}{1+\alpha\mu_i(x)}\right) + (y_i)\ln\left(\frac{\alpha\mu_i(x)}{1+\alpha\mu_i(x)}\right)\Bigg\}. \tag{2.10}$$

We propose, for a random sample of heaped observations, the log-likelihood function (interval-censored regression) is given by the following components

$$\begin{aligned} p_{1i} &= B_I[y_{Li}, \alpha, 1/(1 + (\alpha\mu_i)] \\ p_{2i} &= B_I[y_{Ri} + 1, \alpha, 1/(1 + (\alpha\mu_i)] \end{aligned}$$

where B_I is the three-parameter incomplete beta function. Therefore, under the NB model for heaped data, we have the log-likelihood function

$$\mathcal{L} = \sum_{i=1}^{n} \ln(p_{1i} - p_{2i}) \tag{2.11}$$

Zero-Inflated Models

Sometimes, there exists an excess of zeros in count response data, and Poisson (and other discrete) distribution models may fail in fitting such data. This excess of zeros is called zero-inflation which is shown in falls data (Ullah et al. [2010]), number of defects in manufacturing (Lambert [1992]), number of cubes in the test of tower building for motor development (Cheung [2002]), etc. Due to an increased interest in zero-inflated models, there has been many other studies of statistical analysis using zero-inflated data, Ridout et al. [1998] summarized some literature and cited examples from agriculture, econometrics, patent applications, road safety, species abundance, medical consultations, use of recreational facilities, and even sexual behavior. Hardin and Hilbe [2012] describe the two origins of zero outcomes: individuals who do not enter into the counting process; individuals who enter into the counting process and have a zero outcome. Hence the model must be separated into different parts, one consisting of a zero count $y_i = 0$ and the other consisting of a nonzero count $y_i > 0$.

$$P(Y_i = y_i) = \begin{cases} P_b(y_i = 0) + (1 - P_b(y_i = 0))P_c(y_i = 0) & y_i{=}0 \\ (1 - P_b(y_i = 0))P_c(y_i) & y_i{=}1,2,\ldots \end{cases} \tag{2.12}$$

where P_b is the binary distribution for the probability of a zero outcome, and P_c is the discrete probability function for the count outcomes. Johnson et al. [1992] and Lambert [1992] extensively studied the zero-inflated Poisson distribution (ZIP) which is used in many applications such as econometric counts of purchasing behaviors, counts of sexual behavior episodes, etc. (Ridout et al. [1998]). In our approach, the zero-inflated heaped count data log-likelihood for the Poisson, generalized Poisson, and negative binomial can be specified as

$$\begin{aligned} \mathcal{L} &= \sum_{i \in S} \ln \Big[P_b(y_i = 0) + (1 - P_b(y_i = 0))P_c(y_i = 0) \Big] \\ &\quad + \sum_{i \notin S} \ln \Big[(1 - P_b(y_i = 0))(p_{1i} - p_{2i}) \Big] \end{aligned} \tag{2.13}$$

where S is the set of zero outcomes, $P_b(y_i = 0)$ is the binary model of zero outcomes usually modeled as logistic regression based with specified covariates, and p_{1i} and p_{2i} such that $P_c(y_i) = p_{1i} - p_{2i}$ are as given in the previous sections. The probability of a zero outcome $P_c(y_i = 0)$ are respectively given by $\exp(-\mu_i)$, $\exp(-\mu_i)$, and $\alpha\mu_i/(1 + \alpha\mu_i)^{1+1/\alpha}$ for the Poisson, generalized Poisson, and negative binomial distrbutions.

Therefore, the heaped zero-inflated poisson regression model is

$$
\begin{aligned}
\mathcal{L} &= \sum_{i\in S} \ln\Big[P_b(y_i=0)+(1-P_b(y_i=0))(\exp(-\mu_i)\Big] \\
&\quad + \sum_{i\notin S} \ln\Big[(1-P_b(y_i=0))(\Gamma_I\{y_{Li},\mu_i\}-\Gamma_I\{y_{Ri}+1,\mu_i\})\Big] \qquad (2.14)
\end{aligned}
$$

where P_b is the binary distribution for the probability of a zero outcome. And, the heaped zero-inflated GP regression model is

$$
\begin{aligned}
\mathcal{L} &= \sum_{i\in S} \ln\Big[P_b(y_i=0)+(1-P_b(y_i=0))(\exp(-\mu_i)\Big] \\
&\quad + \sum_{i\notin S} \ln\Big[(1-P_b(y_i=0))(\Gamma_I\{y_{Li}\alpha,\mu_i\}-\Gamma_I\{(y_{Ri}\alpha)+1,\mu_i\})\Big] \qquad (2.15)
\end{aligned}
$$

where P_b is the binary distribution for the probability of a zero outcome. Finally, the heaped zero-inflated NB regression model is

$$
\begin{aligned}
\mathcal{L} &= \sum_{i\in S} \ln\Big[P_b(y_i=0)+(1-P_b(y_i=0))\Big(\frac{\alpha\mu_i}{(1+\alpha\mu_i)^{1+1/\alpha}}\Big)\Big] \\
&\quad + \sum_{i\notin S} \ln\Big[(1-P_b(y_i=0))(B_I[y_{Li},\alpha,1/(1+(\alpha\mu_i)] \\
&\quad - B_I[y_{Ri}+1,\alpha,1/(1+(\alpha\mu_i)])\Big] \qquad (2.16)
\end{aligned}
$$

where P_b is the binary distribution for the probability of a zero outcome.

Chapter 3

Data Analysis

In this chapter, there are four sets of data analysis using our new method of interval-censored regression for heaped count data. In Section 3.1, we present new Stata commands for modeling heaped count data as well as a motivating example using cigarette count data from the National Health and Examination Survey (NHANES 2009-2010) from the Centers for Disease Control and Prevention (CDC). We then illustrate our new regression model, in Section 3.2, using discrete count data from the EBAN study of African American HIV serodiscordant (heterosexual) couples from a National Institute of Mental Health (NIMH) multisite HIV prevention trial. Next, we compare empirical (observed) probabilities, Poisson probabilities, and Heaped Poisson probabilities in a simulation study (see Section 3.3). And finally in Section 3.4, we discuss the derivations of score test statistics for our interval-censored regression models for heaped count data.

3.1 Modeling Heaped Cigarette Count Data

Introduction

In this section, new Stata commands for modeling heaped count data are presented for the Poisson, generalized Poisson, Negative Binomial regression models as well as their Zero-Inflated versions. We illustrate our method of interval-censored regression for heaped count data by analyzing cigarette count data from the National Health and Examination Survey (NHANES 2009-2010) from the Centers for Disease Control and Prevention (CDC) while using the new Stata commands. The Stata commands are implemented through the use of the Stata optimization ml command where we used the lf method. This method requires programs to be written as likelihood functions to allow the software to speed up the computation of numerical derivatives. For the creation of the new Stata commands, we gave Stata our interval-censored regression likelihood functions from Section 2.1 and allowed the software to numerically optimize the derivatives. These commands will be submitted to Stata for public implementation and usage.

Stata Syntax

The accompanying software includes the command files as well as supporting files for prediction and help. In the following syntax diagrams, unspecified *options* include the usual collection of maximization and display options available to all estimation commands. In addition, all zero-inflated commands include the ilink(*linkname*) to specify the link function for the inflation model.

The syntax for specifying a model for heaped count data is given by

```
heapreg depvar [indepvars] [if] [in] [weight] [, exposure(varname_e) offset
    constraints(constraints) vce(vcetype) level(#) irr noheader poisson gpoisson
    nbreg width() heap() hausman]
```

with options poisson, gpoisson, and nbreg for each discrete distribution above, respectively.

While the syntax for heaped zero-inflated count data is given by

```
ziheapreg depvar [indepvars] [if] [in] [weight]
    inflate(varlist[,offset(varname)]| _cons) [, exposure(varname_e)
    constraints(constraints) vce(vcetype) level(#) irr noheader poisson gpoisson
```

`nbreg width() heap() hausman vuong]`

with options poisson, gpoisson, and nbreg for each discrete distribution above, respectively.

A Durbin-Wu-Hausman test, first proposed by Durbin, later modified by Wu and Hausman (Davidson and MacKinnon [1996]) examines to see if there is a significant difference between two models, a more efficient model (heaped) against a less efficient (regular) but consistent model. This occurs to make sure that the more efficient model also gives consistent results. Under the null hypothesis of this test, the estimated coefficients ($\hat{\beta}_p$,$\hat{\beta}_{th}$) are consistent only if $\hat{\beta}_p$ (regular model) is efficient, while under the alternative hypothesis $\hat{\beta}_{th}$ (heaped model) is consistent. Therefore, we have test statistic of

$$a = (\hat{\beta}_p - \hat{\beta}_{th})(V_{th} - V_p)^{-1}(\hat{\beta}_p - \hat{\beta}_{th})^{-1}$$

where V_{th} and V_p are consistent estimates of the covariance matrices of $\hat{\beta}_{th}$ and $\hat{\beta}_p$ respectively. If a significant $p-value$ results, the null hypothesis is rejected therefore meaning that the more efficient model, our heaped version, is better. While, non-significant Hausman test statistic indicate no preference for either model. Results of this test are included in a footnote to the estimation of the model when the user includes the hausman option in any of the commands.

A Vuong test, see (Vuong [1989]), evaluates whether the regression model with zero-inflation or the regression model without zero-inflation is closer to the true model. A random variable ω is defined as the vector $\ln L_Z - \ln L_S$ where L_Z is the likelihood of the zero-inflated model evaluated at its maximum likelihood estimator (MLE) and L_S is the likelihood of the standard (non-zero-inflated) model evaluated at its MLE. The vector of differences over the N observations is then used to define the statistic

$$V = \frac{\sqrt{N}\overline{\omega}}{\sqrt{\sum_i(\omega_i - \overline{\omega})^2/(N-1)}}$$

which, asymptotically, is characterized by a standard normal distribution. A significant positive statistic indicates preference for the zero-inflated model, and a significant negative statistic indicates preference for the model without zero-inflation. Non-significant Vuong statistics indicate no preference for either model. Results of this test are included in a footnote to the estimation of the model when the user includes the vuong option in any of the zero-inflated commands.

Data Analysis: NHANES Example

Using the National Health and Examination Survey (NHANES 2009-2010) data, we model the average number of cigarettes smoked per day during the past 30 days (smd650) as a function of covariates; age (ridageyr), gender (gendernew), and race (racenew), for 1,504 participants. The participants in this study provided informed consent for the collection of data and the data are of de-identified format freely available over the internet (http://www.cdc.gov/nchs/nhanes/nhanes2009-2010/nhanes 09_10.htm, accessed March, 2013). We recoded the original variables ridreth1 variable, now called racenew, that includes Non-Hispanic White versus Others (Mexican American, Other Hispanic, Non-Hispanic Black, Other RaceMulti-Racial) and riagendr. Selected characteristics of the given variables above from the dataset are given in Table 3.3.

To visually investigate where heaping in the average number of cigarettes smoked per day during the past 30 days may exist, we plot the data by the use of a spikeplot in Figure 3.1.

Here we see that heaping tends to be present at multiples of 5 (i.e. 5, 10, 15, etc.). Therefore, we may try the width of heaping of 5 with a half-width of heaping being $\lfloor 5/2 \rfloor$. We also notice that there are no 0's in our outcome variable so the zero-inflated versions of our new commands will not be illustrated in this analysis.

Table 3.1 NHANES Example Selected Characteristics (n = 1504)

Characteristic	Frequency
Cigarettes smoked/day in the past 30 days, mean (SD)	11.55 (9.98)
Age, mean (SD)	40.73 (16.64)
Gender, No. (%)	
Females	669 (44.48)
Males	835 (55.52)
Race, No. (%)	
Non-Hispanic White	749 (49.80)
Other Races	755 (50.20)
Cigarettes smoked/day in the past 30 days, mean (SD)	
Females	11.17 (9.13)
Males	11.85 (10.61)
Non-Hispanic White	14.81 (10.62)
Other Races	8.31 (8.11)

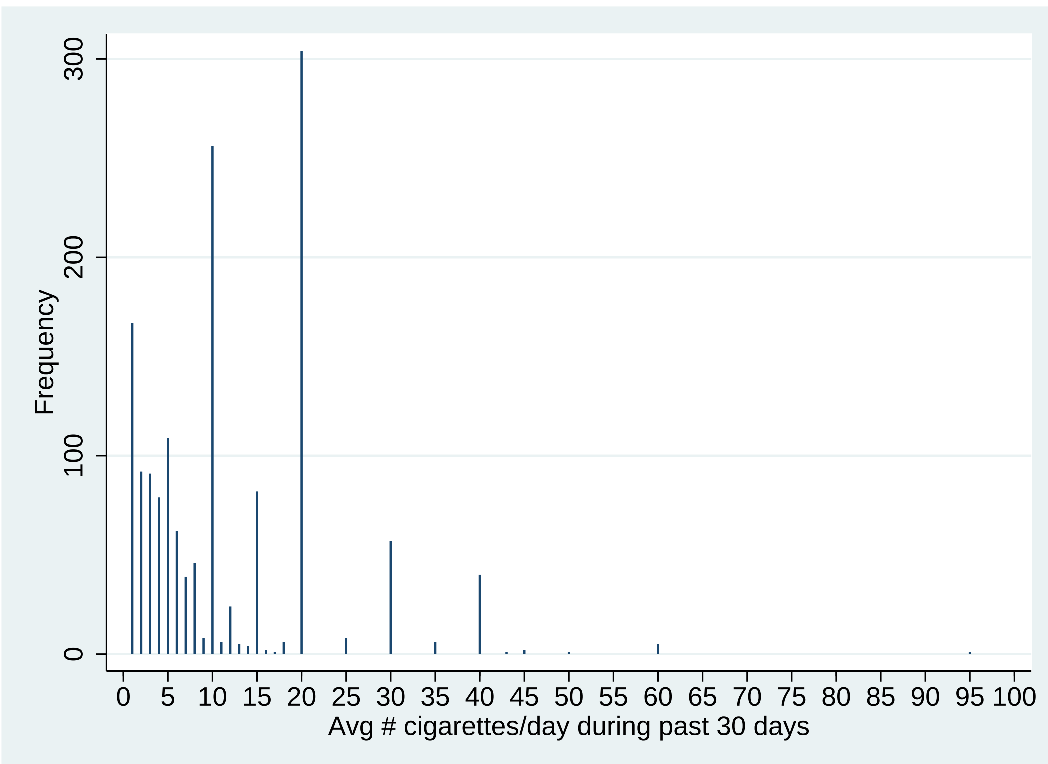

Figure 3.1 Average # of Cigarettes Smoked per day during the past 30 days

Poisson

By fitting a Poisson model (without our proposed approach) to the outcomes, the results are given by

```
. poisson smd650 gendernew racenew ridageyr, nolog

Poisson regression                              Number of obs   =       1504
                                                LR chi2(3)      =    2107.84
                                                Prob > chi2     =     0.0000
Log likelihood = -7782.0546                     Pseudo R2       =     0.1193
```

smd650	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gendernew	-.1100815	.0154432	-7.13	0.000	-.1403497	-.0798134
racenew	.6051288	.0158992	38.06	0.000	.573967	.6362906
ridageyr	.0114867	.0004495	25.56	0.000	.0106057	.0123677
_cons	1.66475	.0246423	67.56	0.000	1.616452	1.713049

Using our proposed method and model, from Section 2.1, to fit the outcomes with heaping at multiples of 5, with a half-width of $\lfloor 5/2 \rfloor$, the results are

```
. heapreg smd650 gendernew racenew ridageyr, width(5) heap(5) poisson hausman nolog

Heaped Poisson regression                       Number of obs   =       1504
Heaping interval(s)   = 5                       LR chi2(3)      =    2052.59
Heaping halfwidth(s) = 2                        Prob > chi2     =     0.0000
Log likelihood = -6199.084                      Pseudo R2       =     0.1420
```

smd650	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gendernew	-.1159769	.0162138	-7.15	0.000	-.1477553	-.0841985
racenew	.6270578	.0167221	37.50	0.000	.5942832	.6598324
ridageyr	.0117447	.0004681	25.09	0.000	.0108271	.0126622
_cons	1.623066	.0257717	62.98	0.000	1.572554	1.673577

```
Hausman test of heaped vs. non-heaped model:       x =    75.62    Pr>x = 0.0000
```

We see a slight difference in the models coefficients and also a statistically significant Hausman test of Heaped Poisson model vs. non-heaped Poisson model.

Generalized Poisson

The results of fitting a GP model (without our proposed approach) to the outcomes are given by

```
. gpoisson smd650 gendernew racenew ridageyr, nolog
```

```
Generalized Poisson regression                  Number of obs   =       1504
                                                LR chi2(3)      =     289.54
Dispersion     =  .6439007                      Prob > chi2     =     0.0000
Log likelihood = -5052.9281                     Pseudo R2       =     0.0279
```

smd650	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gendernew	-.0611733	.0363914	-1.68	0.093	-.1324991	.0101526
racenew	.5461656	.0369862	14.77	0.000	.4736739	.6186573
ridageyr	.0101732	.0009927	10.25	0.000	.0082276	.0121188
_cons	1.738053	.0559818	31.05	0.000	1.628331	1.847775
/atanhdelta	.7648088	.0156008			.7342318	.7953858
delta	.6439007	.0091326			.6256475	.6614493

```
Likelihood-ratio test of delta=0:  chi2(1) = 5458.25       Prob>=chi2 = 0.0000
```

In the regular GP model, we see a statistically significant likelihood-ratio test (LRT) of $\delta = 0$ (dispersion factor), which indicates that the GP model is more appropriate to use than the regular Poisson model. However, by using our proposed method and model, from Section 2.1, to fit the outcomes with heaping at multiples of 5, with a half-width of $\lfloor 5/2 \rfloor$, the results are

```
. heapreg smd650 gendernew racenew ridageyr, width(5) heap(5) gpoisson hausman nolog

Heaped Gen. Poisson regression                  Number of obs   =       1504
Heaping interval(s)  = 5                        LR chi2(4)      =     288.24
Heaping halfwidth(s) = 2                        Prob > chi2     =     0.0000
Log likelihood = -3647.245                      Pseudo R2       =     0.0380
```

smd650	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gendernew	-.0625132	.0368258	-1.70	0.090	-.1346905	.009664
racenew	.5515987	.0374428	14.73	0.000	.4782122	.6249852
ridageyr	.0102635	.0010044	10.22	0.000	.0082949	.0122321
_cons	1.722311	.0566675	30.39	0.000	1.611245	1.833377
/atanhdelta	.7654887	.0157573			.734605	.7963725
delta	.6442986	.0092161			.6258745	.6620039

```
Hausman test of heaped vs. non-heaped model:          x =    15.75    Pr>x = 0.0076
```

Again, we see a slight difference in the coefficients of the models along with a statistically significant Hausman test of Heaped GP model vs. non-heaped GP model.

Negative Binomial

The results of fitting a NB model (without our proposed approach) to the outcomes are given by

```
. nbreg smd650 gendernew racenew ridageyr, nolog

Negative binomial regression                    Number of obs   =       1504
                                                LR chi2(3)      =     290.60
Dispersion     = mean                           Prob > chi2     =     0.0000
Log likelihood = -5048.0101                     Pseudo R2       =     0.0280

------------------------------------------------------------------------------
      smd650 |     Coef.  Std. Err.    z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   gendernew |-.0995121  .0413518  -2.41   0.016     -.1805602   -.0184641
     racenew |  .614582  .0411743  14.93   0.000      .5338819    .6952822
    ridageyr | .0138921  .0013283  10.46   0.000      .0112887    .0164956
       _cons | 1.552952  .0658433  23.59   0.000      1.423901    1.682002
-------------+----------------------------------------------------------------
    /lnalpha |-.6339091  .0425475                    -.7173006   -.5505176
-------------+----------------------------------------------------------------
       alpha | .5305139   .022572                      .488068    .5766512
------------------------------------------------------------------------------
Likelihood-ratio test of alpha=0:  chibar2(01) = 5468.09 Prob>=chibar2 = 0.000
```

In the regular NB model, we see a statistically significant likelihood-ratio test (LRT) of $\alpha = 0$ (dispersion factor), which indicates that the NB model is more appropriate to use than the regular Poisson model. Using our proposed method and model, from Section 2.1, to fit the outcomes with heaping at multiples of 5, with a half-width of $\lfloor 5/2 \rfloor$, the results are

```
. heapreg smd650 gendernew racenew ridageyr, width(5) heap(5) nbreg hausman nolog

Heaped Neg. Binomial regression                 Number of obs   =       1504
Heaping interval(s)   = 5                       LR chi2(4)      =     290.93
Heaping halfwidth(s) = 2                        Prob > chi2     =     0.0000
Log likelihood = -3642.926                      Pseudo R2       =     0.0384

------------------------------------------------------------------------------
      smd650 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

gendernew	-.1019672	.0418746	-2.44	0.015	-.1840399	-.0198944
racenew	.6228999	.0417027	14.94	0.000	.5411642	.7046356
ridageyr	.0140554	.0013436	10.46	0.000	.0114219	.0166889
_cons	1.532817	.0667109	22.98	0.000	1.402066	1.663568
/lnalpha	-.6254518	.0428605			-.7094568	-.5414467
alpha	.5350197	.0229312			.4919113	.5819058

Hausman test of heaped vs. non-heaped model: x =   14.99   Pr>x = 0.0104

Lastly, a slight difference in the models coefficients and dispersion factor (α) is shown and also a statistically significant Hausman test of Heaped Negative Binomial model vs. Negative Binomial model. Our interval regression method for heaped count data, shows to be more efficient than a regular count data model, based on the significance of the Hausman tests for all 3 models with p-values of 0.0104, 0.0000, 0.0076 respectively at $\alpha = 0.05$. All analyses and graphics were preformed using Stata statistical software, version 12 (Stata Corp., College Station, TX).

Discussion

With regard to the reported average number of cigarettes smoked per day during the past 30 days, all variables in our model were statistically significant associated (based on the Heaped NB model) at $\alpha = 0.05$. Females reported fewer average number of cigarettes smoked per day during the past 30 days by a factor of 0.90 (exp(-.110)) compared to males, holding all other factors constant (p-value = 0.015). Non-Hispanic Whites reported more average number of cigarettes smoked per day during the past 30 days by a factor of 1.86 (exp(.623)) compared to other races, holding all other factors constant (p-value $<$ 0.001). Lastly, the reported average number of cigarettes smoked per day during the past 30 days increases by a factor of 1.01 (exp(.014)) as age increases by 1 year, holding all other factors constant (p-value $<$ 0.001). All models have all variables statistically signficantly associated (at $\alpha = 0.05$) with the outcome except GP models (both nonheaped and heaped) where gender is not statistically significant. Our interval regression method for heaped count data, shows to be more efficient than a regular count data model, based on the significance of the Hausman tests for all 3 models with p-values of $0.0104, 0.0000, 0.0076$ respectively at $\alpha = 0.05$.

This section presents a new approach of modeling heaped ("rounded") count data that, by the use of censored interval regression. These heaped count data can lead to biased estimation and imprecision in discrete quantitative data. We also introduce supporting Stata commands and programs, HEAPREG and ZIHEAPREG that illustrate the effectiveness of our approach.

3.2 Modeling Heaped Data with an Application in Self-Reported Frequencies of Sexual Activities

Introduction

The motivation for this section is built around a National Institute of Mental Health (NIMH) multisite HIV (human immunodeficiency virus)/STD (sexually transmitted disease) prevention trial for heterosexual African American couples otherwise known as EBAN trial. The couples were randomized into one of two interventions: the EBAN HIV/STD risk-reduction group (couples) or the EBAN health promotion group (individuals). The goal of this study was to reduce risk behaviors in HIV serodiscordant African American couples (NIMH Multisite HIV/STD Prevention Trial for African American Couples Group [2008]). Researchers observed, in previous literature, that females and males tended to round (heap) their answers to certain questions differently, therefore skewing the final results of the study.

Data Analysis: EBAN Study Data

Data was provided from a National Institute of Mental Health (NIMH) multisite HIV/STD prevention trial for heterosexual African American couples which was conducted in 4 US urban areas: Atlanta (Emory University), Los Angeles (University of California), New York (Columbia University), and Philadelphia (University of Pennsylvania) (Table 3.2). Most participants were from the Columbia University site, followed by the Emory University site, then University of California and University of Pennsylvania sites. The investigators started enrollment in November 2003 and ended in June 2007. The trial was centered around a traditional African concept, where there was a sense of safety, security, and love through the idea of "Fence" (definition of EBAN). This cluster randomized trial used 535 eligible African American HIV serodiscordant heterosexual couples who had the following eligibility criteria:

1. At least 18 years old
2. Be a couple for at least 6 months before study entry
3. Planned to remain a couple at least 12 months after study entry
4. At least 1 partner reported unprotected intercourse in the last 90 days
5. Neither partner has plans to relocate beyond a reasonable distance from the study site
6. At least 1 partner was African American
7. At least one partner agrees that he/she is not planning pregnancy within the next 18 months after study entry
8. Awareness of partner's HIV serostatus
9. Only one partner is HIV seropositive and has known his or her status for at least 3 months before study entry

There were also some exclusion criteria which included the following:

1. One or both partners do not have an address where they can receive mail

2. One or both partners have significant psychiatric, physical, or neurological impairment that would limit their effective participation as confirmed on a Mini Mental State Examination and/or Quick Test

3. History of severe physical or sexual abuse in the 1 year before study entry in the current relationship

4. One or both partners are unwilling or unable to commit to participate in the study through to completion

5. Both partners have previously participated in an HIV sexual risk-reduction intervention for couples in the 12 months before study entry

6. One or both partners are not fluent in English as determined by the informed consent process

Table 3.2 EBAN Study: Randomized Intervention Groups, Overall and by Clinical Site (at Baseline)

Site	Total Participants (#,%)	Total No. in RR Group (#,%)	Total No. in HP Group (#,%)
NY	442 (41.31)	208 (40.00)	234 (42.55)
GA	234 (21.87)	114 (21.92)	120 (21.82)
LA	200 (18.69)	104 (20.00)	96 (17.45)
PA	194 (18.13)	94 (18.08)	100 (18.18)
All Sites	1070 (100)	520 (100)	550 (100)

Couples were randomized into one of two interventions groups: couple-based EBAN HIV/STD risk-reduction (RR) intervention (260 couples) or an individual-based health promotion (HP) comparison (275 couples). In El-Bassel et al. [2010], the authors describe that in the risk-reduction intervention, group sessions addressed community-level factors by emphasizing the threat of HIV to African-American communities. The intervention promotes communication, problem solving, monogamy, and negotiation skills. Some principles that were used to motivate couples to use condoms consistently in order to protech each other and their respective communities include unity, self-determination, and purpose. However, the health promotion comparison group focused on the participants as individuals, not couples. Facilitators discussed behaviors linked to the risk of heart disease, hypertension, stroke, etc. as well as increasing fruit and vegatable consumption, physical activity, medical adherence, and HIV medication adherence.

Some basic characteristics that were recorded were age, education level, monthly income, insured, years lived in the United States, living arrangement, etc. Baseline, immediate post-intervention test (IPT), 6-month and 12-month information was also collected for the following sexual behavior outcomes: Proportion of condom-protected sex, Consistent (100%) condom use, Unprotected sex, and Concurrent partners. The goal of this study was to determine whether the use of a behavioral intervention could reduce the risk of HIV/STD amongst African American HIV serodiscordant (heterosexual) couples (El-Bassel et al. [2010]). For this research, however, we will use the outcome of the question asked to both parties of each couple (respectively): In the *past 90 days*, about how many times did your study partner put his penis into your vagina? and In the *past 90 days*, about how many times did you put your penis into your study partner's vagina? Based on previous literature, we believe males and females may have heaped (rounded) their answers differently.

Data Analysis: Eban Study

To illustrate how the proposed regression models can be applied to real data, we used the EBAN data as discussed in Section 3.2. A selected group of descriptive statistics are described in Table 3.3 by intervention group. The majority of participants, based on the characteristics in Table 3.3, include unemployed, having a high schoold diploma or GED, a monthly income of $400 - 850$, insured, and living with their study partner. Over half of the study participants spent time in an inpatient drug treatment program and about 19% of the couples have concurrent partners. For the following analyses, we used data from the last time period (12 month) and then included the following covariates in a heaped zero-inflated Negative binomial regression model: gender (gender), treatment (trt), partner barriers subscale (xk_pb), and the number of times the participant had sexual intercourse with their partner within the past 90 days at baseline (baseline). Concurrency is important because the outcome variable we analyze is specific to the study partner. That is, each respondent reports the number of episodes of sexual intercourse with the study partner over the past 90 days. Some of the study participants had other (concurrent) sex partners, and any sexual activities with those other partners are not included in our particular outcome. For the inflation (logistic model) part of the model, we specified these same covariates (except baseline) along with age (xage), HIV status at baseline (xhivstatus), concurrent partner (xconcurr), and effect on sexual experience subscale (xk_ese). The subscales (effect on sexual experience and partner barriers) discussed earlier, assess different perceived barriers participants may have towards using condoms. The effect on sexual experience subscale measures perceived aspects of intercourse (sex) that may be perceived as a barrier towards using a condom (i.e. intercouse with a condom is messy). The partner barriers subscale measures perceived partner barriers towards using a condom (i.e my partner controls condom use).

We are interested in the associations of these covariates with the number of reported times having sexual intercourse within the past 90 days after a 12 month follow-up. We determined that responses heaped at multiples of 5 and 12 with respective half-widths of $\lfloor 5/2 \rfloor = 2$ and $\lfloor 12/2 \rfloor = 6$, shown in Figures 3.2 and 3.3. We use the new commands in Stata software (from Section 3.1) to analyze heaped

Table 3.3 EBAN Study: Selected Characteristics for Randomized Intervention Groups (at Baseline)

Characteristic	RR Group (n=520)	HP Group (n=550)
Age, mean (SD)	43.33 (8.00)	43.49 (8.16)
Employment, No. (%)		
Unemployed	369 (71.93)	390 (71.17)
Part-time	45 (8.77)	61 (11.13)
Full-time	99 (19.30)	97 (17.70)
Education, No. (%)		
< Less than a HS diploma	162 (31.52)	164 (29.87)
HS diploma or GED	209 (40.66)	228 (41.53)
Some college/2-year degree	120 (23.35)	136 (24.77)
4-year college degree/post-graduate	23 (4.47)	21 (3.83)
Monthly Income, No. (%)		
< $400	156 (30.41)	151 (27.61)
$400–$850	202 (39.38)	244 (44.61)
$851–$2500	142 (27.68)	137 (25.05)
> $2500	13 (2.53)	15 (2.74)
Insured, No. (%)	377 (73.35)	423 (77.33)
Time spent inpatient drug treatment program, No. (%)	269 (52.33)	285 (52.01)
Living with study partner, No. (%)	368 (71.88)	438 (79.78)
Times in the past 90 days had intercourse, mean (SD)	25.17 (32.87)	23.98 (37.83)
Partner Barriers subscale, mean (SD)	9.98 (3.00)	10.13 (2.95)
Effect on Sexual Experience Subscale, mean (SD)	8.83 (3.31)	8.72 (3.09)
Concurrent partner (by Couple), No (%)	98 (19.14)	98 (18.01)

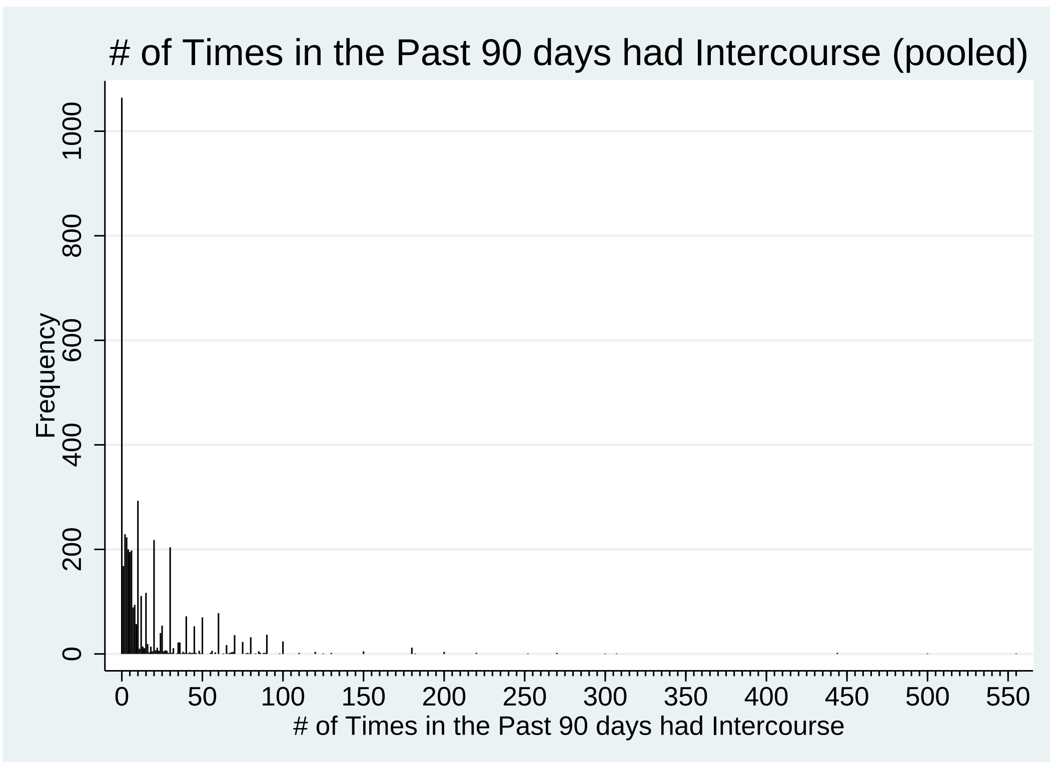

Figure 3.2 EBAN Study: Number of times in the past 90 days had Intercourse (across all 4 time periods)

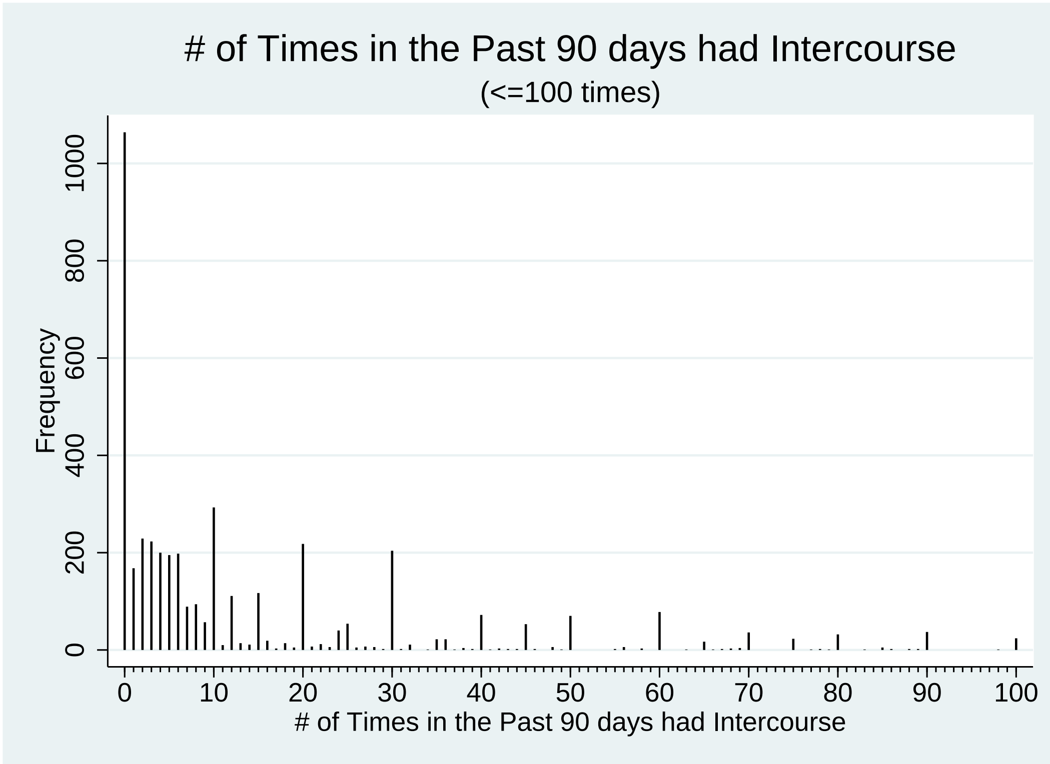

Figure 3.3 EBAN Study: Number of times in the past 90 days had Intercourse less than 100 (across all 4 time periods)

data for the heaped zero-inflated Negative Binomial regression model and the zero-inflated Negative Binomial regression model. Table 3.4 provides effect estimates with associated standard errors, z-values, and p-values for the heaped zero-inflated Negative Binomial regression model, while Table 3.5 provides model estimates from the zero-inflated Negative Binomial regression model where α=dispersion factor.

Overall, there were 936 observations used in this analysis. Here, 253 observations had 0 reported sexual intercourse episodes with their study parnter within the past 90 days and the overall model was significant (p-value = 0.0000) in the heaped zero-inflated Negative Binomial model. With regard to the reported frequency of intercourse within the past 90 days, there were two statistically significant associations. From the results shown in Table 3.4, being in the health promotion (HP) treatment intervention group is associated with fewer reported episodes of sexual intercourse

Table 3.4 Eban Study : Heaped Zero-Inflated NB Estimation Results

Variable	Coefficient	(Std. Err.)	z	P > |z|
Equation 1 : Number of times had sexual intercourse in past 90 days				
trt	-0.183	0.093	-1.96	0.050*
gender	0.019	0.093	0.20	0.839
xk_pb	0.006	0.016	0.38	0.703
baseline	0.017	0.002	9.48	0.000*
Intercept	2.509	0.251	10.01	0.000*
Equation 2 : Inflate				
xage	0.074	0.020	3.70	0.000*
trt	-0.155	0.261	-0.59	0.552
xhivstatus	0.216	0.258	0.84	0.403
gender	0.344	0.287	1.20	0.230
xconcurr	1.668	0.299	5.58	0.000*
xk_ese	-0.098	0.052	-1.89	0.059
xk_pb	0.093	0.058	1.61	0.107
Intercept	-5.892	1.303	-4.52	0.000*
$\ln(\alpha)$	0.377	0.084		
α	1.458	0.122		

*p-value < 0.05

Table 3.5 Eban Study : Zero-Inflated NB Estimation Results

Variable	Coefficient	(Std. Err.)	z	P > |z|
Equation 1 : Number of times had sexual intercourse in past 90 days				
trt	-0.179	0.092	-1.94	0.052
gender	0.015	0.092	0.16	0.871
xk_pb	0.006	0.016	0.38	0.703
baseline	0.016	0.002	9.50	0.000*
Intercept	2.529	0.248	10.21	0.000*
Equation 2 : Inflate				
xage	0.072	0.019	3.73	0.000*
trt	-0.144	0.253	-0.57	0.569
xhivstatus	0.219	0.251	0.87	0.383
gender	0.330	0.278	1.19	0.236
xconcurr	1.631	0.290	5.63	0.000*
xk_ese	-0.095	0.050	-1.91	0.056
xk_pb	0.093	0.056	1.65	0.099
Intercept	-5.756	1.258	-4.58	0.000*
$\ln(\alpha)$	0.352	0.083		
α	1.422	0.118		

*p-value < 0.05

by a factor of 0.83 ($\exp(-0.183)$) compared to the risk-reduction (RR) intervention group, holding all other factors constant (p-value = 0.050). As the number of reported episodes of sexual intercourse at baseline increases by one year, the odds of reporting episodes of sexual intercourse increases by a factor of 1.02 ($\exp(0.017)$), holding all other variables constant (p-value < 0.001. Being female is associated, but not statistically significant, with more reported episodes of sexual intercourse by a factor of 1.02 ($\exp(0.019)$) compared to males, holding all other variables constant (p-value = 0.839). As a person's partner barriers subscale increases by one, the odds of reporting episodes of sexual intercourse increases, but not significantly, by a factor of 1.01 ($\exp(0.006)$), holding all other variables constant (p-value = 0.703.

From the inflation equation, from Table 3.4, that is a result from a logistic regression model predicting a reported frequency of sexual intercourse of 0, includes two statistically significant assocations as well. As age increases by 10 years, the odds of reporting zero episodes of sexual intercourse increases by a factor of 2.10 ($\exp(0.074 * 10)$), holding all other variables constant (p-value< 0.001). Having a concurrent partner increases the odds of reporting zero episodes of sexual intercourse by a factor of 5.30 ($\exp(1.668)$) compared to not having a concurrent partner, holding all other variables constant (p-value < 0.001). The nonsignificant associations where the odds of reporting zero episodes of sexual intercourse increased includes being HIV+, being female, and the partner barriers subscale. While the nonsignificant associations where the odds of reporting zero episodes of sexual intercourse decreased includes treatment intervention group and the effect on sexual experience subscale.

Finally, we point out that a zero-inflated negative binomial regression model, shown in Table 3.5, which does not take into account heaping results in a nonsignificant association of the treatment intervention groups. Though this would not always happen, it does highlight the benefit of properly addressing the distribution of the outcomes. Other than the difference in declaring the association of group membership significant, the inference from the model which does not address heaping was the same. All analyses were performed using Stata statistical software, version 12 (Stata Corp., College Station, TX).

Discussion

We have developed statistical models for heaped count data where our method introduces a mixture of likelihood functions for heaped and nonheaped count data. For the heaped count data, we considered the reported outcome to be censored over the half width of the heaping multiple. We then simultaneously consider nonheaped count data where we treat the data as exact counts and base them on the distribution's likelihood. This proposed method was motivated by self-reported frequency of sexual intercourse from the EBAN study of African American HIV serodiscordant African American couples. Due to the nature of our self-reported study data, we noticed clear heaping for the number of times each study participant had sexual intercourse with their study partner within the past 90 days at multiples of 5 and 12, which may be the result of recall or measurement errors.

The analysis of this data using a heaped zero-inflated negative binomial regression model having a significant treatment intervention effect, reveals a possible advantage of using a heaped model rather than a non-heaped model. Gender, however, does not substantially effect the expected the number of times of having sexual intercourse with the study partner, within the past 90 days. Being in the treatment group and the number of sexual intercourse episodes at baseline are all associated with reporting significantly fewer episodes and greater episodes of sexual intercourse, respectively. We again point out that these reported numbers are the number of episodes of sexual intercourse with the partner in this study. Thus, persons with concurrent partners may have greater numbers of sexual intercourse episodes, just not greater with the partner in this study. Having a concurrent partner increased the odds of reporting zero episodes of sexual intercourse significantly. Accounting for heaping improved the fit of the count data, while preserving the exact self-reported counts. Our method requires some prior knowledge of the multiple(s) of heaping (or half-width of heaping) which some researchers may not have, therefore further research is necessary to address these issues.

3.3 Simulation Study

Introduction

In this section, we compare Poisson probabilities, Heaped Poisson probabilities, and the empirical (observed) probabilities for each particular covariate pattern. The model used in our simulations was defined by

$$y = 1 + x_1 + 2x_2 - 0x_3 \tag{3.1}$$

where we synthesize x_1 from a Bernoulli(0.5) distribution, x_2 from a Bernoulli(0.5) distribution, and x_3 from a Bernoulli(0.5) distribution. However, we heaped the outcome data, y, based on multiples of 4 while using a Poisson distribution (from Section 2.1).

Data Analysis: Simulation Study

The spikeplot of our simulated data is shown in Figure 3.4. For each covariate coefficient in Equation 3.1, the Poisson, Heaped Poisson, and empirical probabilities were computed. Simulation results are based on 10000 replications. Some properties of the heaped Poisson data include

$$E(Y_i) = 21.1934 \qquad Var(Y_i) = 437.1595$$
$$Min(Y_i) = 0 \qquad Max(Y_i) = 80$$

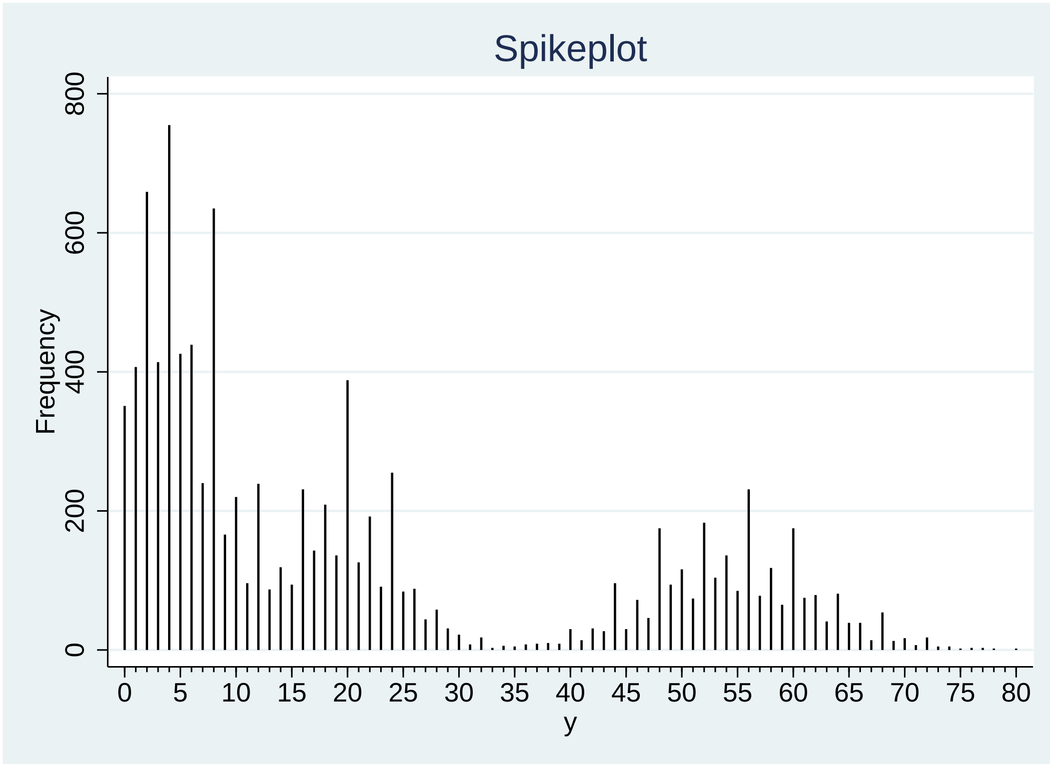

Figure 3.4 Simulation Study: Spikeplot of Heaped Poisson data (10,000 Replications)

Table 3.6 and 3.7 provides effect estimates with associated standard errors, z-values, and p-values from Poisson and Heaped Poisson distributions respectively.

Table 3.6 Simulation Study: Poisson Regression Estimation Results

Poisson Regression				
Variable	**Coefficient**	**(Std. Err.)**	**z**	**P > |z|**
x1	1.004	(0.005)	204.20	0.000*
x2	2.013	(0.007)	297.80	0.000*
x3	-0.0004	(0.004)	-0.09	0.928
Intercept	0.983	(0.008)	129.52	0.000*

*p-value < 0.05

Both models were overall statistically significant (p-value < 0.0001). We also performed a Hausman test for these data. The Hausman test (Davidson and MacKinnon [1996]) examines to see if there is a significant difference between two models, a more

Table 3.7 Simulation Study: Heaped Poisson Regression Estimation Results

Heaped Poisson Regression				
Variable	**Coefficient**	**(Std. Err.)**	**z**	**P > |z|**
x1	1.011	(0.005)	200.98	0.000*
x2	2.034	(0.007)	284.82	0.000*
x3	-0.0006	(0.004)	-0.150	0.884
Intercept	0.956	(0.008)	119.94	0.000*

*p-value < 0.05

efficient model (heaped) against a less efficient (regular) but consistent model. This occurs to make sure that the more efficient model also gives consistent results. Under the null hypothesis of this test, the estimated coefficients ($\hat{\beta}_p$,$\hat{\beta}_{th}$) are consistent only if $\hat{\beta}_p$ (regular model) is efficient, while under the alternative hypothesis $\hat{\beta}_{th}$ (heaped model) is consistent. Therefore, we have test statistic of

$$a = (\hat{\beta}_p - \hat{\beta}_{th})(V_{th} - V_p)^{-1}(\hat{\beta}_p - \hat{\beta}_{th})^{-1}$$

where V_{th} and V_p are consistent estimates of the covariance matrices of $\hat{\beta}_{th}$ and $\hat{\beta}_p$ respectively. If a significant p-value results, the null hypothesis is rejected therefore meaning that the more efficient model, our heaped version, is better. While, non-significant Hausman test statistic indicate no preference for either model. Results of this test concluded in a test statistic of 133.20 with a p-value < 0.0001, therefore our heaped model is the more efficient model. The means from each covariate pattern from their respective distributions (see Table 3.8).

The probabilities from each regression model by the 8 covariate patterns for x_1, x_2, and x_3 are shown in Table 3.9-3.17 and visually represented in Figure 3.5.

Each covariate pattern has a different set of probabilities based on the means from Table 3.8. The empirical probabilities (in black) show the heaped poisson simulated data, while the Poisson regression model probabilities are in red. Visually, the bar charts show that the Poisson regression model does not fit the heaped data accurately due to the black bars being so far from the red (Poisson) distribution. In Figure 3.6,

Table 3.8 Simulation Study: Means from using the Empirical, Poisson, Heaped Poisson Distribution regression models

			Means		
x_1	x_2	x_3	Empirical	Poisson	Heaped Poisson
0	0	0	2.6315	2.6720	2.6026
0	0	1	2.5883	2.6709	2.6009
1	0	0	7.3539	7.2955	7.1565
1	0	1	7.3603	7.2926	7.1519
0	1	0	20.0629	20.0040	19.9022
0	1	1	20.0630	19.9962	19.8895
1	1	0	54.5407	54.6185	54.7275
1	1	1	54.5503	54.5972	54.6924

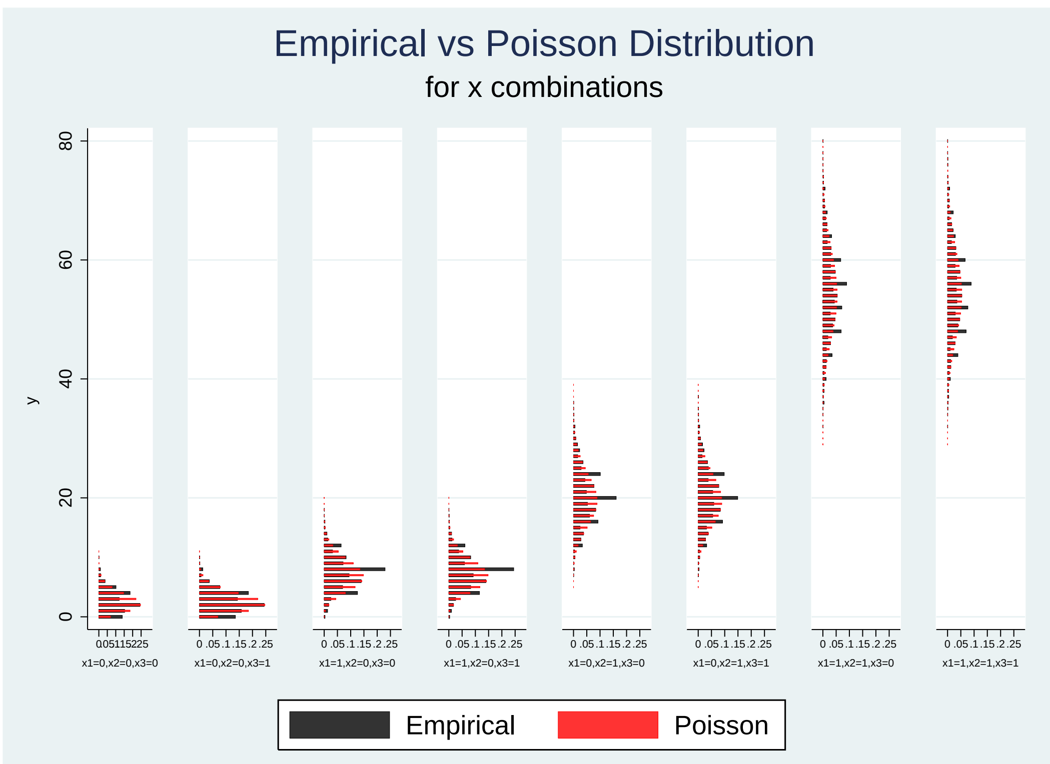

Figure 3.5 Simulation Study: All Covariate Patterns Probabilities

we have extracted just the covariate combination of $x_1 = 1$, $x_2 = 1$, and $x_3 = 1$ from Figure 3.5 to magnify the insufficient Poisson model fitting heaped count data.

Notice, in Figure 3.7, we have extracted the same covariate combination of $x_1 = 1$, $x_2 = 1$, and $x_3 = 1$ to magnify the more efficient Heaped Poisson model fitting heaped

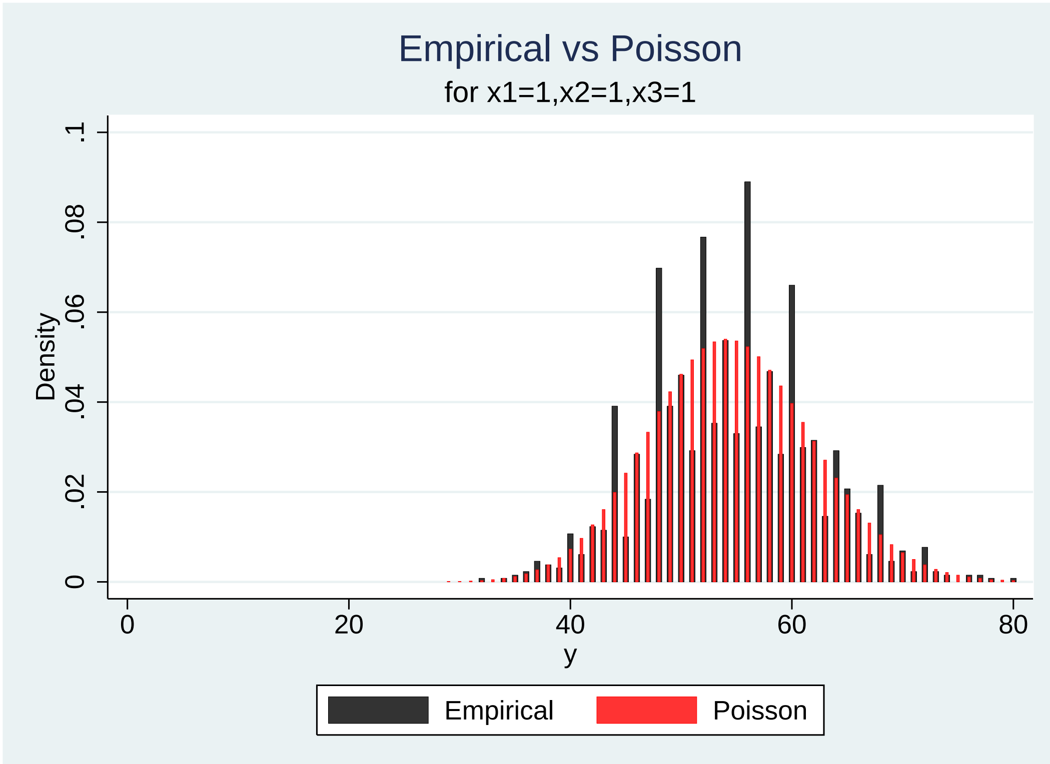

Figure 3.6 Simulation Study: Covariate Pattern $x_1 = 1$, $x_2 = 1$, and $x_3 = 1$ probabilities

count data better than a regular Poisson regression model.

The probabilities for the true parameter estimates (from Equation 3.1) using the Empirical, Poisson, and Heaped Poisson regression models with 10,000 replications are located in Tables 3.9-3.17. The probabilities for the true parameter estimates using the Observed Censored and Heaped Censored Poisson regression models, at multiples of 4 are located in Tables 3.18, 3.19, 3.20, 3.21. The graphs associated with the other 7 combinations comparing the Observed Censored to the Heaped Censored Poisson models is located in Figures 3.8-3.14. All analyses and graphics were preformed using Stata statistical software, version 12 (Stata Corp., College Station, TX).

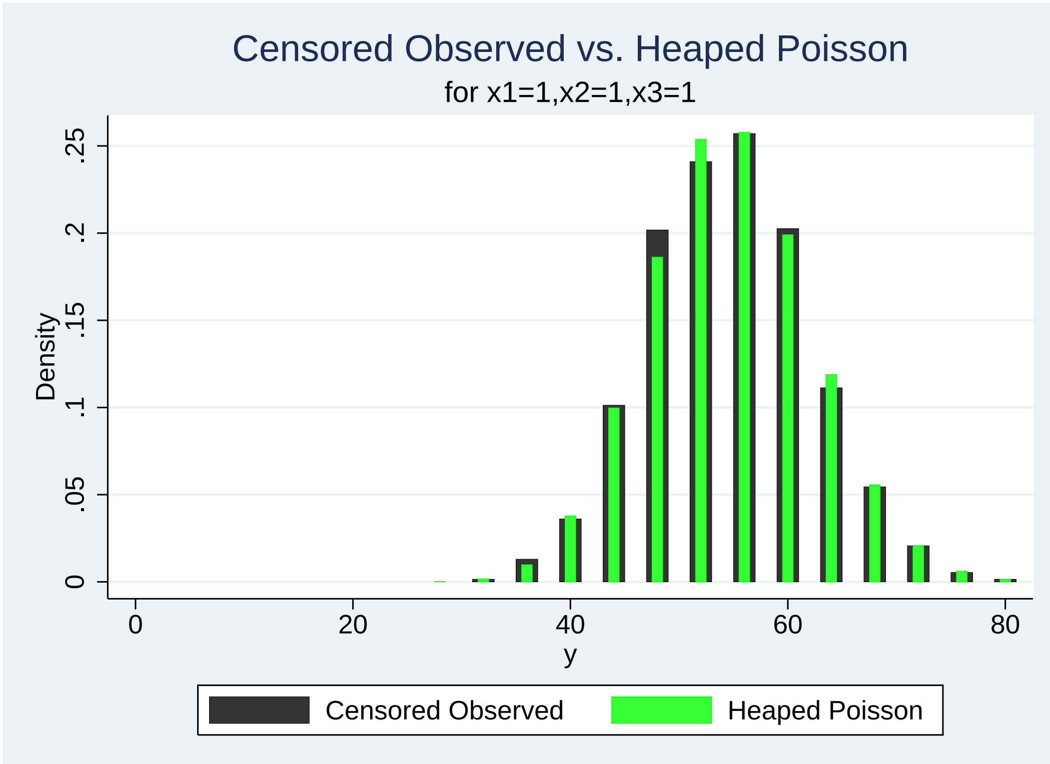

Figure 3.7 Simulation Study: Covariate Pattern $x_1 = 1$, $x_2 = 1$, and $x_3 = 1$ Censored probabilities

Discussion

We have developed statistical models that handle heaped count data. This method introduces a mixture of likelihood functions for heaped and nonheaped count data where heaped data is assumed to be censored observations over an interval or constant multiple. The nonheaped data we treat as exact counts from it's respective distribution. We illustrated the effectiveness of our new approach by preforming a simulation study where we synthesized heaped data and fit it both with a Poisson regression model and Heaped Poisson regression model. Based on the results of this study, our heaped Poisson model fits the heaped count data better than a regular Poisson regression model.

Table 3.9 Simulation Study: Probabilities (10000 replications) for the true parameter estimates (from Equation 3.1) using the Empirical, Poisson, and Heaped Poisson regression models for x_1=0, x_2=0, x_3=0 & x_1=0, x_2=0, x_3=1 & x_1=1, x_2=0, x_3=0 (part A)

	Covariate Pattern								
	x_1=0,x_2=0,x_3=0			x_1=0,x_2=0,x_3=1			x_1=1,x_2=0,x_3=0		
y	Emp	P	HP	Emp	P	HP	Emp	P	HP
0	0.1386	0.0691	0.0741	0.1354	0.0692	0.0742	0.0032	0.0007	0.0008
1	0.1528	0.1847	0.1928	0.1576	0.1848	0.193	0.013	0.005	0.0056
2	0.2433	0.2467	0.2509	0.2439	0.2468	0.251	0.017	0.0181	0.02
3	0.1205	0.2197	0.2177	0.1433	0.2197	0.2176	0.0252	0.0439	0.0476
4	0.185	0.1468	0.1416	0.1845	0.1467	0.1415	0.1258	0.0801	0.0852
5	0.1016	0.0784	0.0737	0.0768	0.0784	0.0736	0.0698	0.1169	0.122
6	0.037	0.0349	0.032	0.0372	0.0349	0.0319	0.1396	0.1421	0.1455
7	0.011	0.0133	0.0119	0.0063	0.0133	0.0119	0.0942	0.1481	0.1487
8	0.0094	0.0045	0.0039	0.0127	0.0044	0.0039	0.2297	0.1351	0.1331
9	0	0.0013	0.0011	0.0016	0.0013	0.0011	0.0714	0.1095	0.1058
10	0.0008	0.0004	0.0003	0.0008	0.0004	0.0003	0.0828	0.0799	0.0757
11	0	0.0001	0.0001	0	0.0001	0.0001	0.0317	0.053	0.0493
12	0	0	0	0	0	0	0.0641	0.0322	0.0294
13	0	0	0	0	0	0	0.0138	0.0181	0.0162
14	0	0	0	0	0	0	0.0106	0.0094	0.0083
15	0	0	0	0	0	0	0.0032	0.0046	0.0039
16	0	0	0	0	0	0	0.0032	0.0021	0.0018
17	0	0	0	0	0	0	0.0008	0.0009	0.0007
18	0	0	0	0	0	0	0.0008	0.0004	0.0003
19	0	0	0	0	0	0	0	0.0001	0.0001
20	0	0	0	0	0	0	0	0.0001	0
21	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0
33	0	0	0	0	0	0	0	0	0
34	0	0	0	0	0	0	0	0	0
35	0	0	0	0	0	0	0	0	0
36	0	0	0	0	0	0	0	0	0
37	0	0	0	0	0	0	0	0	0

Table 3.10 Simulation Study: Probabilities (10000 replications) for the true parameter estimates (from Equation 3.1) using the Empirical, Poisson, and Heaped Poisson regression models for x_1=0, x_2=0, x_3=0 & x_1=0, x_2=0, x_3=1 & x_1=1, x_2=0, x_3=0 (part B)

	Covariate Pattern								
	x_1=0,x_2=0,x_3=0			x_1=0,x_2=0,x_3=1			x_1=1,x_2=0,x_3=0		
y	Emp	P	HP	Emp	P	HP	Emp	P	HP
38	0	0	0	0	0	0	0	0	0
39	0	0	0	0	0	0	0	0	0
40	0	0	0	0	0	0	0	0	0
41	0	0	0	0	0	0	0	0	0
42	0	0	0	0	0	0	0	0	0
43	0	0	0	0	0	0	0	0	0
44	0	0	0	0	0	0	0	0	0
45	0	0	0	0	0	0	0	0	0
46	0	0	0	0	0	0	0	0	0
47	0	0	0	0	0	0	0	0	0
48	0	0	0	0	0	0	0	0	0
49	0	0	0	0	0	0	0	0	0
50	0	0	0	0	0	0	0	0	0
51	0	0	0	0	0	0	0	0	0
52	0	0	0	0	0	0	0	0	0
53	0	0	0	0	0	0	0	0	0
54	0	0	0	0	0	0	0	0	0
55	0	0	0	0	0	0	0	0	0
56	0	0	0	0	0	0	0	0	0
57	0	0	0	0	0	0	0	0	0
58	0	0	0	0	0	0	0	0	0
59	0	0	0	0	0	0	0	0	0
60	0	0	0	0	0	0	0	0	0
61	0	0	0	0	0	0	0	0	0
62	0	0	0	0	0	0	0	0	0
63	0	0	0	0	0	0	0	0	0
64	0	0	0	0	0	0	0	0	0
65	0	0	0	0	0	0	0	0	0
66	0	0	0	0	0	0	0	0	0
67	0	0	0	0	0	0	0	0	0
68	0	0	0	0	0	0	0	0	0
69	0	0	0	0	0	0	0	0	0
70	0	0	0	0	0	0	0	0	0
71	0	0	0	0	0	0	0	0	0
72	0	0	0	0	0	0	0	0	0

Table 3.11 Simulation Study: Probabilities (10000 replications) for the true parameter estimates (from Equation 3.1) using the Empirical, Poisson, and Heaped Poisson regression models for x_1=0, x_2=0, x_3=0 & x_1=0, x_2=0, x_3=1 & x_1=1, x_2=0, x_3=0 (part C)

	Covariate Pattern								
	x_1=0,x_2=0,x_3=0			x_1=0,x_2=0,x_3=1			x_1=1,x_2=0,x_3=0		
y	Emp	P	HP	Emp	P	HP	Emp	P	HP
73	0	0	0	0	0	0	0	0	0
74	0	0	0	0	0	0	0	0	0
75	0	0	0	0	0	0	0	0	0
76	0	0	0	0	0	0	0	0	0
77	0	0	0	0	0	0	0	0	0
78	0	0	0	0	0	0	0	0	0
79	0	0	0	0	0	0	0	0	0
80	0	0	0	0	0	0	0	0	0

Table 3.12 Simulation Study: Probabilities (10000 replications) for the true parameter estimates (from Equation 3.1) using the Empirical, Poisson, and Heaped Poisson regression models for x_1=1, x_2=0, x_3=1 & x_1=0, x_2=1, x_3=0 & x_1=0, x_2=1, x_3=1 (part A)

	Covariate Pattern								
	x_1=1,x_2=0,x_3=1			x_1=0,x_2=1,x_3=0			x_1=0,x_2=1,x_3=1		
y	Emp	P	HP	Emp	P	HP	Emp	P	HP
0	0.0032	0.0007	0.0008	0	0	0	0	0	0
1	0.0097	0.005	0.0056	0	0	0	0	0	0
2	0.017	0.0181	0.02	0	0	0	0	0	0
3	0.0258	0.044	0.0478	0	0	0	0	0	0
4	0.1155	0.0802	0.0854	0	0	0	0	0	0
5	0.0824	0.117	0.1221	0	0.0001	0.0001	0	0.0001	0.0001
6	0.1397	0.1422	0.1456	0	0.0002	0.0002	0	0.0002	0.0002
7	0.0921	0.1481	0.1488	0.0008	0.0005	0.0006	0.0008	0.0005	0.0006
8	0.2447	0.135	0.133	0.0032	0.0013	0.0014	0.0033	0.0013	0.0014
9	0.059	0.1094	0.1057	0.0008	0.0029	0.0031	0.0008	0.0029	0.0031
10	0.0824	0.0798	0.0756	0.0056	0.0058	0.0061	0.0057	0.0058	0.0061
11	0.0372	0.0529	0.0491	0.004	0.0106	0.011	0.0049	0.0106	0.0111
12	0.0606	0.0321	0.0293	0.0335	0.0176	0.0183	0.0319	0.0177	0.0184
13	0.0121	0.018	0.0161	0.0279	0.0271	0.0281	0.0278	0.0272	0.0282
14	0.0105	0.0094	0.0082	0.0375	0.0387	0.0399	0.0376	0.0388	0.04
15	0.0032	0.0046	0.0039	0.0247	0.0516	0.0529	0.0311	0.0517	0.0531
16	0.0024	0.0021	0.0018	0.0924	0.0645	0.0658	0.0917	0.0646	0.066
17	0.0016	0.0009	0.0007	0.0606	0.0759	0.0771	0.0548	0.076	0.0772
18	0.0008	0.0004	0.0003	0.0837	0.0844	0.0852	0.0835	0.0844	0.0853
19	0	0.0001	0.0001	0.0526	0.0888	0.0893	0.0589	0.0889	0.0893
20	0	0.0001	0	0.161	0.0888	0.0888	0.1489	0.0888	0.0888
21	0	0	0	0.0486	0.0846	0.0842	0.0548	0.0846	0.0841
22	0	0	0	0.0773	0.0769	0.0761	0.0777	0.0769	0.076
23	0	0	0	0.043	0.0669	0.0659	0.0376	0.0668	0.0658
24	0	0	0	0.1012	0.0558	0.0546	0.0982	0.0557	0.0545
25	0	0	0	0.0287	0.0446	0.0435	0.0385	0.0445	0.0434
26	0	0	0	0.0359	0.0343	0.0333	0.0352	0.0343	0.0332
27	0	0	0	0.0159	0.0254	0.0245	0.0147	0.0254	0.0244
28	0	0	0	0.0231	0.0182	0.0174	0.0221	0.0181	0.0174
29	0	0	0	0.0151	0.0125	0.012	0.0164	0.0125	0.0119
30	0	0	0	0.0088	0.0084	0.0079	0.009	0.0083	0.0079
31	0	0	0	0.004	0.0054	0.0051	0.0033	0.0054	0.0051
32	0	0	0	0.0056	0.0034	0.0032	0.0057	0.0034	0.0031
33	0	0	0	0.0016	0.002	0.0019	0.0016	0.002	0.0019
34	0	0	0	0.0016	0.0012	0.0011	0.0016	0.0012	0.0011
35	0	0	0	0.0008	0.0007	0.0006	0.0008	0.0007	0.0006
36	0	0	0	0.0008	0.0004	0.0004	0	0.0004	0.0003
37	0	0	0	0	0.0002	0.0002	0.0008	0.0002	0.0002

Table 3.13 Simulation Study: Probabilities (10000 replications) for the true parameter estimates (from Equation 3.1) using the Empirical, Poisson, and Heaped Poisson regression models for x_1=1, x_2=0, x_3=1 & x_1=0, x_2=1, x_3=0 & x_1=0, x_2=1, x_3=1 (part B)

	Covariate Pattern								
	x_1=1,x_2=0,x_3=1			x_1=0,x_2=1,x_3=0			x_1=0,x_2=1,x_3=1		
y	Emp	P	HP	Emp	P	HP	Emp	P	HP
38	0	0	0	0	0.0001	0.0001	0	0.0001	0.0001
39	0	0	0	0	0.0001	0.0001	0	0.0001	0
40	0	0	0	0	0	0	0	0	0
41	0	0	0	0	0	0	0	0	0
42	0	0	0	0	0	0	0	0	0
43	0	0	0	0	0	0	0	0	0
44	0	0	0	0	0	0	0	0	0
45	0	0	0	0	0	0	0	0	0
46	0	0	0	0	0	0	0	0	0
47	0	0	0	0	0	0	0	0	0
48	0	0	0	0	0	0	0	0	0
49	0	0	0	0	0	0	0	0	0
50	0	0	0	0	0	0	0	0	0
51	0	0	0	0	0	0	0	0	0
52	0	0	0	0	0	0	0	0	0
53	0	0	0	0	0	0	0	0	0
54	0	0	0	0	0	0	0	0	0
55	0	0	0	0	0	0	0	0	0
56	0	0	0	0	0	0	0	0	0
57	0	0	0	0	0	0	0	0	0
58	0	0	0	0	0	0	0	0	0
59	0	0	0	0	0	0	0	0	0
60	0	0	0	0	0	0	0	0	0
61	0	0	0	0	0	0	0	0	0
62	0	0	0	0	0	0	0	0	0
63	0	0	0	0	0	0	0	0	0
64	0	0	0	0	0	0	0	0	0
65	0	0	0	0	0	0	0	0	0
66	0	0	0	0	0	0	0	0	0
67	0	0	0	0	0	0	0	0	0
68	0	0	0	0	0	0	0	0	0
69	0	0	0	0	0	0	0	0	0
70	0	0	0	0	0	0	0	0	0
71	0	0	0	0	0	0	0	0	0
72	0	0	0	0	0	0	0	0	0

Table 3.14 Simulation Study: Probabilities (10000 replications) for the true parameter estimates (from Equation 3.1) using the Empirical, Poisson, and Heaped Poisson regression models for x_1=1, x_2=0, x_3=1 & x_1=0, x_2=1, x_3=0 & x_1=0, x_2=1, x_3=1 (part C)

	Covariate Pattern								
	x_1=1,x_2=0,x_3=1			x_1=0,x_2=1,x_3=0			x_1=0,x_2=1,x_3=1		
y	Emp	P	HP	Emp	P	HP	Emp	P	HP
73	0	0	0	0	0	0	0	0	0
74	0	0	0	0	0	0	0	0	0
75	0	0	0	0	0	0	0	0	0
76	0	0	0	0	0	0	0	0	0
77	0	0	0	0	0	0	0	0	0
78	0	0	0	0	0	0	0	0	0
79	0	0	0	0	0	0	0	0	0
80	0	0	0	0	0	0	0	0	0

Table 3.15 Simulation Study: Probabilities (10000 replications) for the true parameter estimates (from Equation 3.1) using the Empirical, Poisson, and Heaped Poisson regression models for x_1=1, x_2=1, x_3=0 & x_1=1, x_2=1, x_3=1 (part A)

	Covariate Pattern					
	x_1=1,x_2=1,x_3=0			x_1=1,x_2=1,x_3=1		
y	Emp	P	HP	Emp	P	HP
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0
7	0	0	0	0	0	0
8	0	0	0	0	0	0
9	0	0	0	0	0	0
10	0	0	0	0	0	0
11	0	0	0	0	0	0
12	0	0	0	0	0	0
13	0	0	0	0	0	0
14	0	0	0	0	0	0
15	0	0	0	0	0	0
16	0	0	0	0	0	0
17	0	0	0	0	0	0
18	0	0	0	0	0	0
19	0	0	0	0	0	0
20	0	0	0	0	0	0
21	0	0	0	0	0	0
22	0	0	0	0	0	0
23	0	0	0	0	0	0
24	0	0	0	0	0	0
25	0	0	0	0	0	0
26	0	0	0	0	0	0
27	0	0	0	0	0	0
28	0	0	0	0	0	0
29	0	0.0001	0	0	0.0001	0.0001
30	0	0.0001	0.0001	0	0.0001	0.0001
31	0	0.0002	0.0002	0	0.0002	0.0002
32	0.0008	0.0003	0.0003	0.0008	0.0003	0.0003
33	0	0.0005	0.0005	0	0.0005	0.0005
34	0.0008	0.0008	0.0007	0.0008	0.0008	0.0007
35	0.0008	0.0012	0.0011	0.0015	0.0012	0.0011
36	0.0041	0.0018	0.0017	0.0023	0.0018	0.0017

Table 3.16 Simulation Study: Probabilities (10000 replications) for the true parameter estimates (from Equation 3.1) using the Empirical, Poisson, and Heaped Poisson regression models for x_1=1, x_2=1, x_3=0 & x_1=1, x_2=1, x_3=1 (part B)

	Covariate Pattern					
	x_1=1,x_2=1,x_3=0			x_1=1,x_2=1,x_3=1		
y	Emp	P	HP	Emp	P	HP
37	0.0016	0.0026	0.0026	0.0046	0.0027	0.0026
38	0.0041	0.0038	0.0037	0.0038	0.0038	0.0037
39	0.0041	0.0053	0.0052	0.0031	0.0054	0.0052
40	0.0123	0.0073	0.0071	0.0107	0.0073	0.0071
41	0.0049	0.0097	0.0094	0.0061	0.0097	0.0095
42	0.0123	0.0126	0.0123	0.0123	0.0127	0.0124
43	0.0131	0.016	0.0156	0.0115	0.0161	0.0157
44	0.0345	0.0199	0.0194	0.0391	0.0199	0.0196
45	0.0131	0.0241	0.0236	0.01	0.0242	0.0238
46	0.0288	0.0286	0.0281	0.0284	0.0287	0.0283
47	0.0189	0.0332	0.0327	0.0184	0.0333	0.0329
48	0.069	0.0378	0.0373	0.0698	0.0379	0.0375
49	0.037	0.0422	0.0417	0.0391	0.0423	0.0419
50	0.046	0.0461	0.0456	0.046	0.0462	0.0458
51	0.0279	0.0493	0.049	0.0292	0.0494	0.0491
52	0.0715	0.0518	0.0515	0.0767	0.0519	0.0516
53	0.0435	0.0534	0.0532	0.0353	0.0534	0.0533
54	0.0542	0.054	0.0539	0.0537	0.054	0.054
55	0.0386	0.0536	0.0537	0.033	0.0536	0.0537
56	0.0896	0.0523	0.0525	0.089	0.0523	0.0524
57	0.0279	0.0501	0.0504	0.0345	0.0501	0.0503
58	0.0468	0.0472	0.0475	0.0468	0.0471	0.0474
59	0.0288	0.0437	0.0441	0.0284	0.0436	0.044
60	0.0674	0.0398	0.0402	0.066	0.0397	0.0401
61	0.0296	0.0356	0.0361	0.0299	0.0355	0.0359
62	0.0312	0.0314	0.0318	0.0315	0.0313	0.0317
63	0.0164	0.0272	0.0277	0.0146	0.0271	0.0275
64	0.0329	0.0232	0.0237	0.0292	0.0231	0.0235
65	0.014	0.0195	0.0199	0.0207	0.0194	0.0198
66	0.0156	0.0161	0.0165	0.0153	0.0161	0.0164
67	0.0099	0.0132	0.0135	0.0061	0.0131	0.0134
68	0.0164	0.0106	0.0109	0.0215	0.0105	0.0108
69	0.0058	0.0084	0.0086	0.0046	0.0083	0.0085
70	0.0066	0.0065	0.0067	0.0069	0.0065	0.0067
71	0.0016	0.005	0.0052	0.0023	0.005	0.0051
72	0.0082	0.0038	0.0039	0.0077	0.0038	0.0039

Table 3.17 Simulation Study: Probabilities (10000 replications) for the true parameter estimates (from Equation 3.1) using the Empirical, Poisson, and Heaped Poisson regression models for x_1=1, x_2=1, x_3=0 & x_1=1, x_2=1, x_3=1 (part C)

	Covariate Pattern					
	x_1=1,x_2=1,x_3=0			x_1=1,x_2=1,x_3=1		
y	Emp	P	HP	Emp	P	HP
73	0.0025	0.0029	0.003	0.0023	0.0028	0.0029
74	0.0025	0.0021	0.0022	0.0015	0.0021	0.0022
75	0.0008	0.0015	0.0016	0	0.0015	0.0016
76	0.0008	0.0011	0.0011	0.0015	0.0011	0.0011
77	0.0008	0.0008	0.0008	0.0015	0.0008	0.0008
78	0.0008	0.0005	0.0006	0.0008	0.0005	0.0006
79	0	0.0004	0.0004	0	0.0004	0.0004
80	0.0008	0.0003	0.0003	0.0008	0.0003	0.0003

Table 3.18 Simulation Study: Probabilities (10000 replications) for the true parameter estimates (shown in Figures 3.8-3.9) using the Observed Censored and Heaped Censored Poisson regression models

	Covariate Pattern			
	x_1=0,x_2=0,x_3=0		x_1=0,x_2=0,x_3=1	
y	Obs Censored	Heaped Poisson	Obs Censored	Heaped Poisson
0	0.5347	0.5178	0.5369	0.5182
4	0.6874	0.7159	0.6857	0.7156
8	0.0582	0.0492	0.0586	0.0491
12	0.0008	0.0004	0.0008	0.0004
16	0	0	0	0
20	0	0	0	0
24	0	0	0	0
28	0	0	0	0
32	0	0	0	0
36	0	0	0	0
40	0	0	0	0
44	0	0	0	0
48	0	0	0	0
52	0	0	0	0
56	0	0	0	0
60	0	0	0	0
64	0	0	0	0
68	0	0	0	0
72	0	0	0	0
76	0	0	0	0
80	0	0	0	0

Table 3.19 Simulation Study: Probabilities (10000 replications) for the true parameter estimates (shown in Figures 3.10-3.11) using the Observed Censored and Heaped Censored Poisson regression models

	Covariate Pattern			
	x_1=1,x_2=0,x_3=0		x_1=1,x_2=0,x_3=1	
y	Obs Censored	Heaped Poisson	Obs Censored	Heaped Poisson
0	0.0332	0.0264	0.0299	0.0264
4	0.3774	0.4203	0.3804	0.4209
8	0.6177	0.6088	0.6179	0.6087
12	0.203	0.1789	0.2028	0.1783
16	0.0186	0.015	0.0185	0.0149
20	0.0008	0.0004	0.0008	0.0004
24	0	0	0	0
28	0	0	0	0
32	0	0	0	0
36	0	0	0	0
40	0	0	0	0
44	0	0	0	0
48	0	0	0	0
52	0	0	0	0
56	0	0	0	0
60	0	0	0	0
64	0	0	0	0
68	0	0	0	0
72	0	0	0	0
76	0	0	0	0
80	0	0	0	0

Table 3.20 Simulation Study: Probabilities (10000 replications) for the true parameter estimates (shown in Figures 3.12-3.13) using the Observed Censored and Heaped Censored Poisson regression models

	Covariate Pattern			
	x_1=0,x_2=1,x_3=0		x_1=0,x_2=1,x_3=1	
y	Obs Censored	Heaped Poisson	Obs Censored	Heaped Poisson
0	0	0	0	0
4	0	0.0003	0	0.0003
8	0.0104	0.0114	0.0106	0.0114
12	0.1085	0.1034	0.1079	0.1038
16	0.2989	0.3209	0.2987	0.3216
20	0.4232	0.4236	0.4238	0.4235
24	0.2861	0.2734	0.2872	0.2729
28	0.0988	0.0951	0.0974	0.0948
32	0.0216	0.0192	0.0212	0.0191
36	0.0032	0.0024	0.0032	0.0023
40	0	0.0002	0	0.0001
44	0	0	0	0
48	0	0	0	0
52	0	0	0	0
56	0	0	0	0
60	0	0	0	0
64	0	0	0	0
68	0	0	0	0
72	0	0	0	0
76	0	0	0	0
80	0	0	0	0

Table 3.21 Simulation Study: Probabilities (10000 replications) for the true parameter estimates (shown in Figures 3.14 & 3.7) using the Observed Censored and Heaped Censored Poisson regression models

	Covariate Pattern			
	x_1=1,x_2=1,x_3=0		x_1=1,x_2=1,x_3=1	
y	Obs Censored	Heaped Poisson	Obs Censored	Heaped Poisson
0	0	0	0	0
4	0	0	0	0
8	0	0	0	0
12	0	0	0	0
16	0	0	0	0
20	0	0	0	0
24	0	0	0	0
28	0	0.0001	0	0.0002
32	0.0016	0.0018	0.0016	0.0018
36	0.0114	0.0098	0.013	0.0098
40	0.0377	0.0377	0.036	0.0379
44	0.1018	0.099	0.1013	0.0998
48	0.1997	0.1854	0.2017	0.1864
52	0.2431	0.2532	0.2409	0.2538
56	0.2571	0.258	0.257	0.2578
60	0.2038	0.1997	0.2026	0.1991
64	0.1101	0.1196	0.1113	0.1189
68	0.0543	0.0562	0.0544	0.0558
72	0.0214	0.021	0.0207	0.0208
76	0.0057	0.0063	0.0053	0.0063
80	0.0016	0.0013	0.0016	0.0013

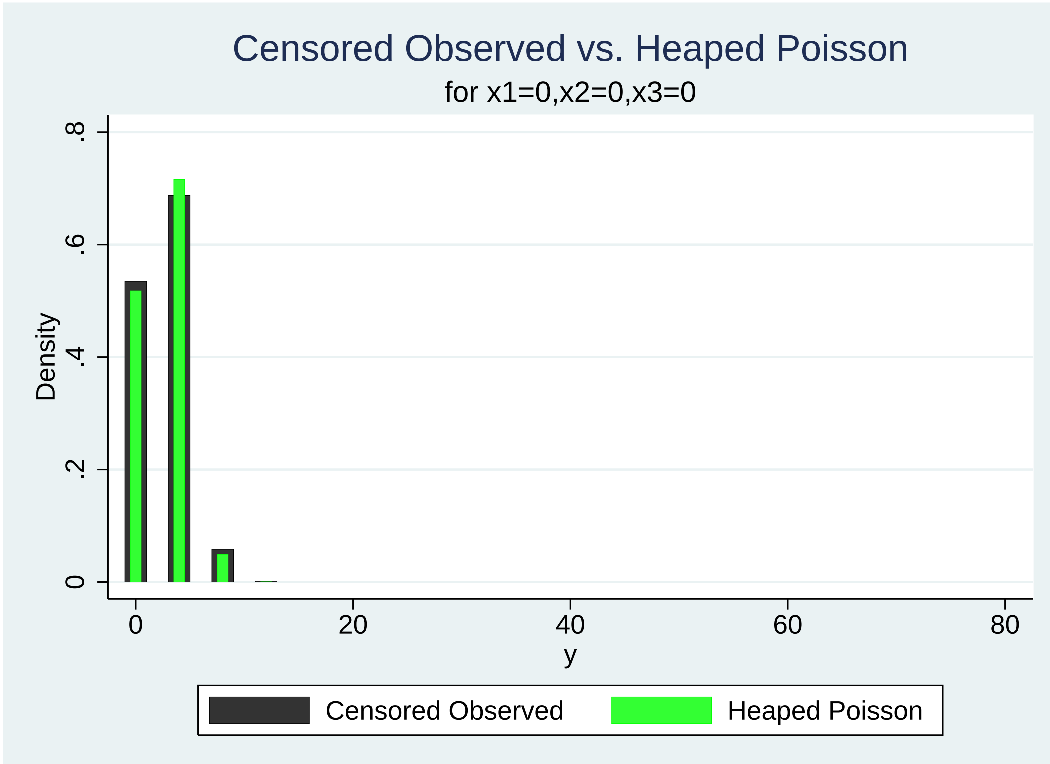

Figure 3.8 Simulation Study: Covariate Pattern $x_1 = 0$, $x_2 = 0$, and $x_3 = 0$ Censored probabilities

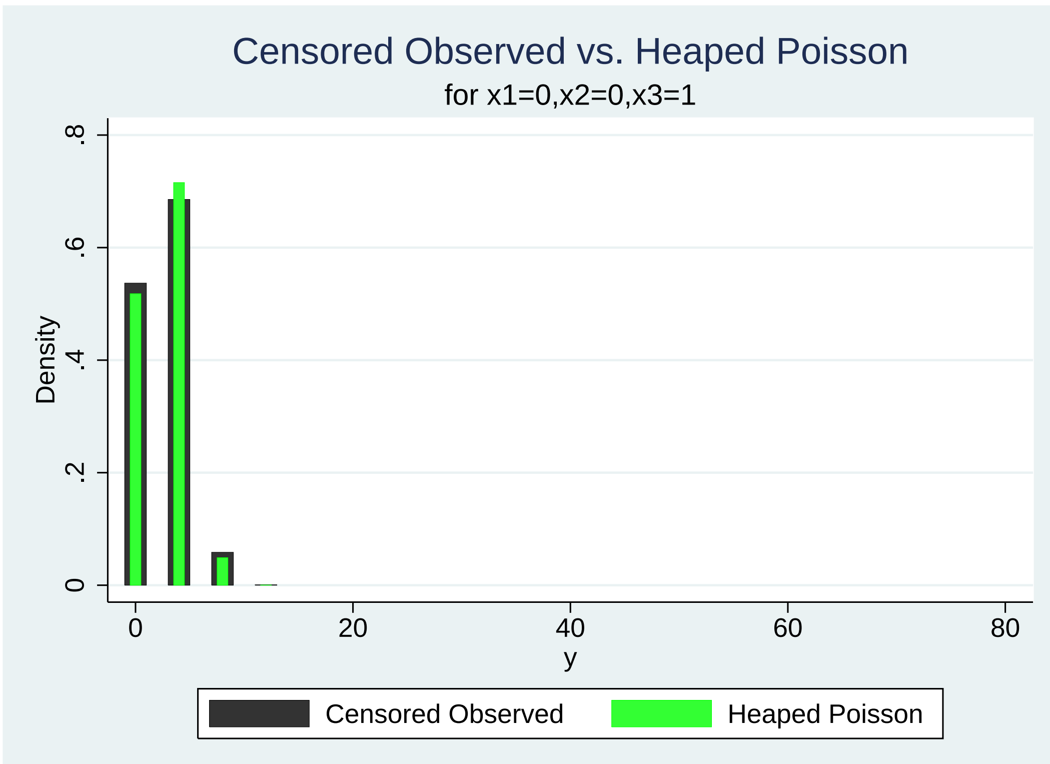

Figure 3.9 Simulation Study: Covariate Pattern $x_1 = 0$, $x_2 = 0$, and $x_3 = 1$ Censored probabilities

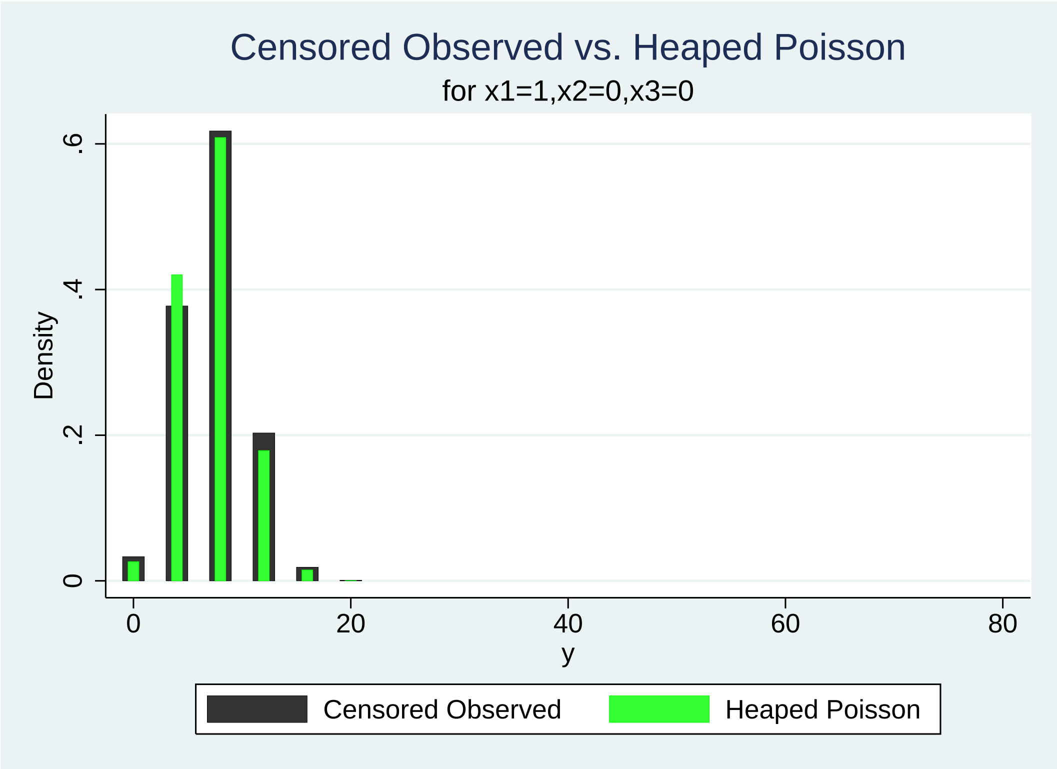

Figure 3.10 Simulation Study: Covariate Pattern $x_1 = 1$, $x_2 = 0$, and $x_3 = 0$ Censored probabilities

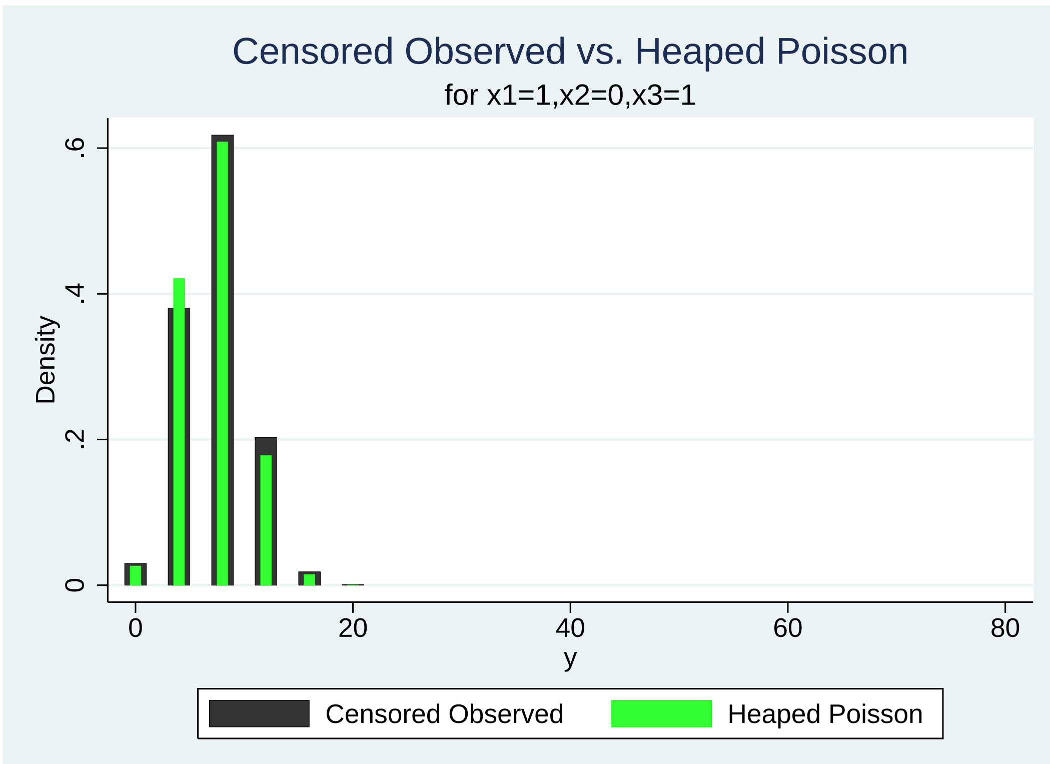

Figure 3.11 Simulation Study: Covariate Pattern $x_1 = 1$, $x_2 = 0$, and $x_3 = 1$ Censored probabilities

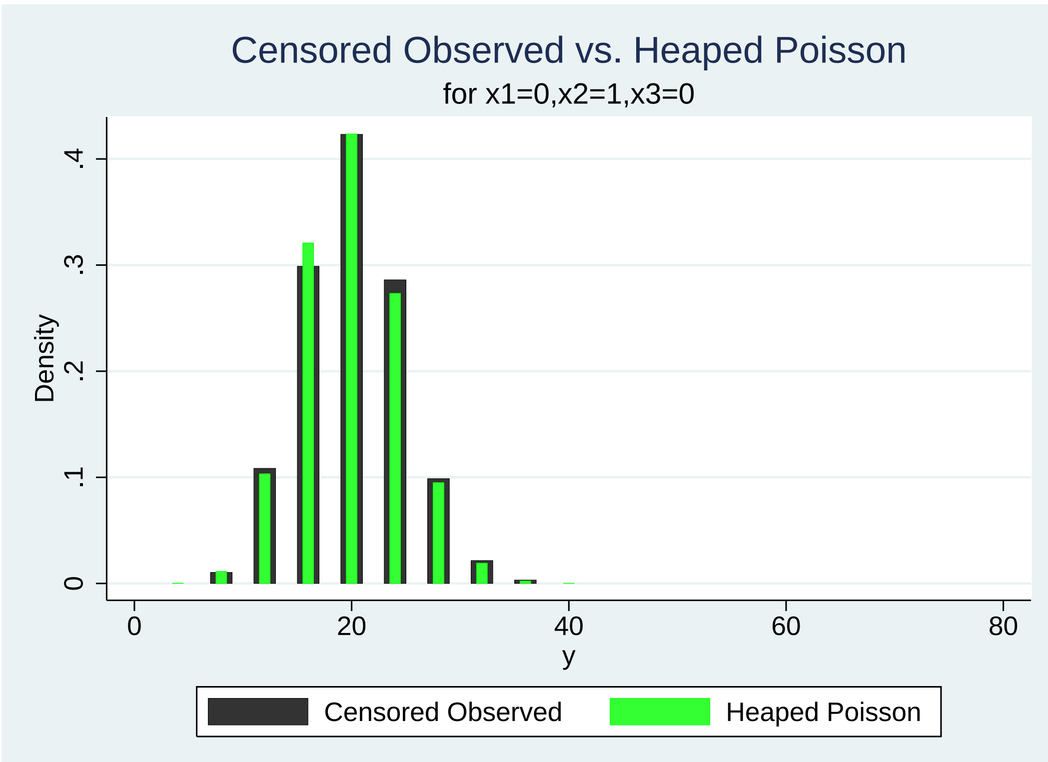

Figure 3.12 Simulation Study: Covariate Pattern $x_1 = 0$, $x_2 = 1$, and $x_3 = 0$ Censored probabilities

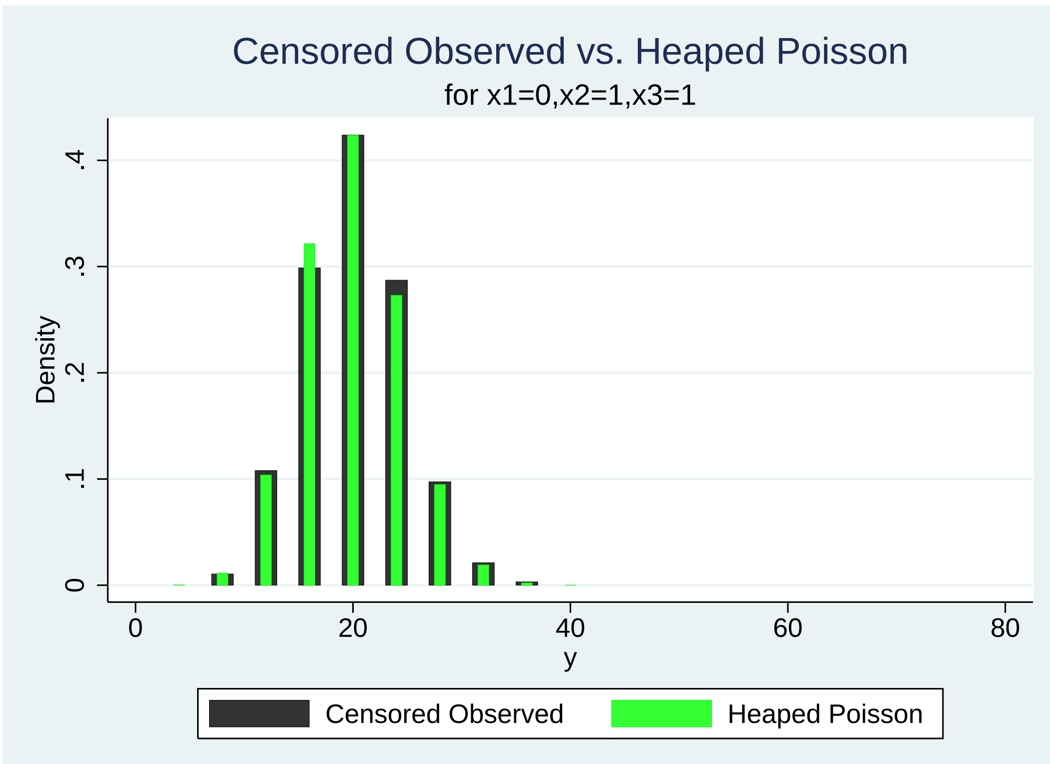

Figure 3.13 Simulation Study: Covariate Pattern $x_1 = 0$, $x_2 = 1$, and $x_3 = 1$ Censored probabilities

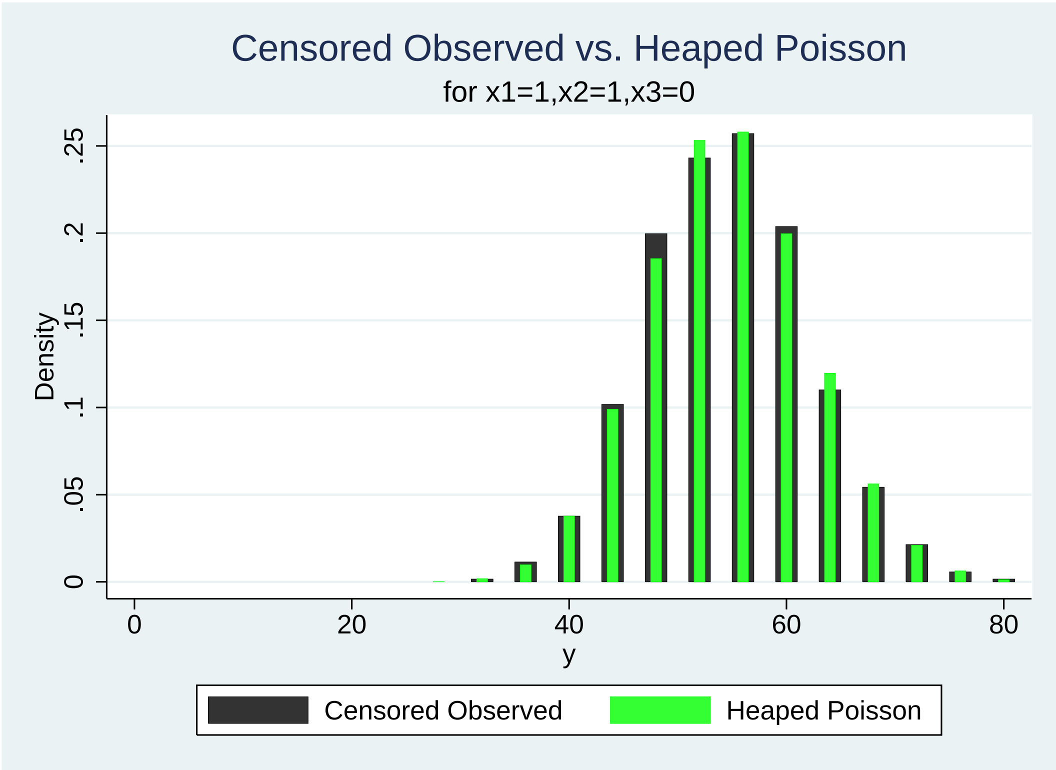

Figure 3.14 Simulation Study: Covariate Pattern $x_1 = 1$, $x_2 = 1$, and $x_3 = 0$ Censored probabilities

3.4 Score Test Derivatives for Overdispersion in Heaped Count Data Models

In this section, we will discuss score test derivations for overdispersion in heaped count data models. Instead of computing both model when performing a likelihood-ratio test (LRT), or computing the alternative model only and performing a Wald test, the score test avoids the computation of the alternative model altogether. We have developed the first derivatives of our interval-censored regression models to compute a score test for heaped count regression models.

Score Test Derivatives

Several tests have been proposed to determine the amount of overdispersion in the Poisson model (Cameron and Trivedi [1986]; Dean and Lawless [1989]). A score test for overdispersion derived by Yang et al. [2009] based on the GP-2 model is given by

$$\mathcal{S}(\hat{\beta}) = \frac{1}{2n}\Big(\sum_{i=1}^{n}\Big(\frac{y_i(y_i-1)}{\hat{\mu}} - y_i\Big)\Big)^2$$

which is a χ_1^2. Another score test for overdispersion in Poisson model based on the NB regression model was derived by Cameron and Trivedi [1986] and Dean [1992] is given by

$$\mathcal{S}(\hat{\beta}) = \Big(\frac{\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 - y_i}{\sqrt{2\sum_{i=1}^{n}\hat{\mu}_i^{\,2}}}\Big)^2$$

where $\hat{\mu}_i$ is the predicted count under the Poisson model.

The score test is also referred to as the Lagrange Multiplier test or the Rao test. It is the most powerful test when the true value of the parameter is close to the null value. The main advantage is that its calculation only requires evaluation under the null hypothesis. In the case of overdispersion tests, this means that one need only evaluate a Poisson regression model. Using those results, a test of overdispersion can then be calculated versus other models which allow overdispersion via a dispersion parameter α. The test in Poisson regression models is carried out for comparing $H_0 : \alpha = 0$ versus $H_1 : \alpha > 0$.

It is an extension of the Fisher dispersion test and was formulated from the Taylor expansion of the loglikelihood. Therefore, for any regression model for which there is a vector of regression parameters and an additional parameter, we can derive a score test of that additional dispersion parameter. An advantage of this test is that a model for which the additional parameter does not need to be estimated. The maximum likelihood estimation (MLE) of the regression parameters is augmented with zero for the scalar (α), and that augmented vector is used to evaluate the terms of the test statistic.

Let the loglikelihood function of the unrestricted model be $\mathcal{L}(\theta)$ where θ is the the augmented parameter vector comprised of β, α. The first derivative of the log-likelihood is written in terms of the partitioned vector as

$$\begin{aligned} \mathcal{U}(\theta^T) &= \left(\frac{\partial \mathcal{L}}{\partial \theta^T}\right)_{1\times(p+1)} \\ &= \begin{bmatrix} \dfrac{\partial \mathcal{L}}{\partial \beta^T_{1\times p}} & \dfrac{\partial \mathcal{L}}{\partial \alpha^T_{1\times 1}} \end{bmatrix} \end{aligned} \tag{3.2}$$

where $\mathcal{U}(\theta^T)$ is called the partial score vector. The matrix of second derivatives in terms of the covariates and associated diagonal weight terms are

$$\left(\frac{\partial^2 \mathcal{L}}{\partial \theta \partial \theta^T}\right)_{(p+1)\times(p+1)} = \begin{bmatrix} \dfrac{\partial^2 \mathcal{L}}{\partial \beta \partial \beta^T_{p\times p}} & \dfrac{\partial^2 \mathcal{L}}{\partial \beta \partial \alpha^T_{p\times 1}} \\ \\ \dfrac{\partial^2 \mathcal{L}}{\partial \alpha \partial \beta^T_{1\times p}} & \dfrac{\partial^2 \mathcal{L}}{\partial \alpha \partial \alpha^T_{1\times 1}} \end{bmatrix} \tag{3.3}$$

This is helpful when estimating the variance through the use of the expected value or Fisher information matrix given as

$$
\begin{aligned}
-E\begin{bmatrix} \dfrac{\partial^2\mathcal{L}}{\partial\beta\partial\beta^T} & \dfrac{\partial^2\mathcal{L}}{\partial\beta\partial\alpha^T} \\ \\ \dfrac{\partial^2\mathcal{L}}{\partial\alpha\partial\beta^T} & \dfrac{\partial^2\mathcal{L}}{\partial\alpha\partial\alpha^T} \end{bmatrix} &= \mathcal{J}(\beta,\alpha) \\
&= \begin{bmatrix} \mathcal{A}(\beta,\alpha) & \mathcal{C}(\beta,\alpha) \\ \mathcal{C}(\beta,\alpha)^T & \mathcal{B}(\beta,\alpha) \end{bmatrix} \qquad (3.4)
\end{aligned}
$$

The inverse of this matrix gives the asymptotic variance of the maximum likelihood estimate. The estimated variance of $\mathcal{U}_\alpha(\hat{\beta},0)$ is the element of the inverse of this matrix,

$$\mathcal{B}*(\beta,\alpha) = (\mathcal{B}(\beta,\alpha) - \mathcal{C}(\beta,\alpha)^T[\mathcal{A}(\beta,\alpha)]^{-1}\mathcal{C}(\beta,\alpha))^{-1} \qquad (3.5)$$

which is $\mathcal{J}(\beta,\alpha)^{-1}$ corresponding to α. The score test is then given by

$$\mathcal{S} = [\mathcal{U}_\alpha(\hat{\beta},0)]^T\mathcal{B}*(\beta,0)[\mathcal{U}_\alpha(\hat{\beta},0)] \qquad (3.6)$$

where $S \sim \chi^2_q$ and q is the dimension of α. The first derivatives of the heaped regression models and heaped zero-inflated regression models are shown in Appendices A and B, respectively. The second derivatives for each regression model are extremely complicated and not necessary to calculate a score test.

Discussion

In literature, score tests are often preferred over LRT and Wald tests due to not having to compute both models when performing a LRT, or computing the alternative model only performing a Wald test. Software companies are increasingly starting to include score test estimates in their software. However, for this research the score test for heaped overdispersion in regression models has extremely complicated analytic derivatives and numerical derivatives may be easier to compute.

Chapter 4

Conclusion

4.1 Summary

This dissertation develops a new method to analyze heaped count data that results when subjects recall the frequency of events prefer for reporting from a limited set of rounded responses or preferred digits over reporting exact counts. These rounded responses and digit preferences (also referred to as data coarsening) could be characterized by reported frequencies (or counts) favoring multiples of 20, reporting counts ending with 0 or 5, or a preference for reporting an even number over an odd number. This mixture of values is a type of measurement error (pattern of misreporting) that can lead to biased estimation and imprecision in discrete qua ntitative data. Sometimes this pattern in data can be explained or understood, but its effect on the statistical inference may be harder to anticipate. A visual representation of heaped data can be seen in a frequency distribution (histogram) where the h eaps are represented as periodic peaks or spikes within the overall data layout.

We proposed statistical models to model heaped count data using a mixture of likelihood functions for heaped and nonheaped count data. We also created new heaped count data regression commands in Stata statistical software where we considered the reporte d outcome is actually censored over the half width of the heaping multiple for heaped count data. We also considered that nonheaped (not censored) data follow the count distribution's likelihood for exact counts. The investigator would need to specify th e heaping multiples over which heaped values are censored via an interval regression approach for our new method. We illustrated our new method and Stata commands for handling heaped count data with two real-world data applications and one simulation st udy. The average number of cigarettes smoked per day during the past 30 days as a function of age, gender, and race for 1,504 participants from NHANES data was studied where we saw heaping at multiples of 5 (half-width of $\lfloor 5/2 \rfloor$). We sho wed that by using our interval-censored regression method, based on the Poisson, GP, and NB models, the heaped versions are more efficient than the regular versions based on the significant results from the Hausman tests. Then we investigated self-repor ted frequency of sexual intercourse from the EBAN study of African American HIV serodiscordant heterosexual couples. We noticed clear heaping based on the spikeplot of the number of times the participant had sexual intercourse with their study partner w ithin the past 90 days. Heaping was present at multiples of 5 (half-width of $\lfloor 5/2 \rfloor$) and 12 (half-width of $\lfloor 12/2 \rfloor$), which may have been a result of recall or measurement errors. By modeling this data using the heaped zero- inflated NB regression model may have had an advantage over the regular zero-inflated NB regression model due to the significant treatment intervention effect. For the simulation study, we illustrated the effectiveness of our new approach by synthesizin g heaped Poisson data and fitting the data based on the Poisson regression model and heaped Poisson regression model. We also compared the empirical (observed), Poisson, and heaped Poisson probabilities and based on these probabilities and graphs, as we ll as the results (Hausman test) from the Poisson and heaped Poisson regression models, we conclude that the heaped Poisson regression model was a better fit. Finally, we derived score test derivatives for our interval-censored regression models for hea ped data and concluded that these models have extremely complicated analytic derivatives and numerical derivatives may be more appropriate to use.

Some advantages of new method include the following: Non-Bayesian approach, censorship (probabilistic) covers the heaping interval (half-width), considers all multiples of heaping, not certain ending digits, use of statistical models and predictions (no Whipple's index or Myers' blended index), no 'set' heaping mechanism, and no multiple imputation. However, there are some limitations based on our interval-censored regression method. One limitation is that our method doesn't explain when heaping occurs or who does it and if the investigator specifies too many heaping multiples, that may create numerical problem in the model convergence.

4.2 Future Work

The purpose of this section is to propose potential future work, which will extend the materials presented in this dissertation. We plan to explore a method that incorporates a modeling component which explains heaping in count data occur and who does it. In our current proposed method of interval-censored regression, the censored model converts those heaped values into an interval of possible outcomes under the assumption that the reported value is actually a multiple of a frequency from a smaller scale.

In this new approach, we will be able to simultaneously model using another set of covariates the likelihood of reporting a heaped value to see when heaping in count data occur and what are the characteristics of those reporters. In this approach, reported values on heaping multiples are treated as a mixture of exact reports and interval reports (assumed to be scaled up from a smaller period of time).

Finally, we can include our interval-censored regression method for heaped count data with other discrete distributions. In future developments, we will address the generalized negative binomial, zero-inflated generalized negative binomial, Poisson-inverse gaussian, as well as zero-truncated versions of these distributions.

Bibliography

A. Colin Cameron and Pravin K. Trivedi. Econometric models based on count data: Comparisons and applications of some estimators and tests. *Journal of Applied Econometrics*, 1:29–53, 1986.

Andrew A. R. Channon, Sabu S. Padmadas, and John W. McDonald. Measuring birth weight in developing countries: Does the method of reporting in retrospective surveys matter? *Maternal and Child Health Journal*, 15:12–18, 2011.

Y. B. Cheung. Zero-inflated models for regression analysis of count data: a study of growth and development. *Statistics in Medicine*, 21:1461–1469, 2002.

P. C. Consul. *Generalized Poisson Distributions: Properties and Applications*. Marcel Dekker, New York, 1989.

P. C. Consul and F. Famoye. Generalized poisson regression model. *Communications in Statistics—Theory and Methods*, 21:89–109, 1992.

Russell Davidson and James G. MacKinnon. Specification tests based on artificial regressions. *Journal of the American Statistical Association*, 85:220–227, 1996.

C. Dean and J. F. Lawless. Tests for detecting overdispersion in poisson regression models. *Journal of the American Statistical Association*, 84:467–472, 1989.

C. B. Dean. Testing for overdispersion in poisson and binomial regression models. *JASA*, 87:451–457, 1992.

Srdjan Denic, Falah Khatib, and Hussein Saadi. Quality of age data in patients from developing countries. *Journal of Public Health*, 26:168–171, 2004.

Nabila El-Bassel, John B. Jemmott, J. Richard Landis, Willo Pequegnat, Gina M. Wingood, Gail E. Wyatt, and Scarlett L. Bellamy. National institute of mental health multisite eban hiv/std prevention intervention for african american hiv serodiscordant couples. *ARCH INTERN MED*, 170:1594–1601, 2010.

Peter Foldvari, Bas Van Leeuwen, and Jieli Van Leeuwen-Li. How did women count? a note on gender-specific age heaping differences in the sixteenth to nineteenth centuries. *Economic History Review*, 65:304–313, 2012.

J. W. Hardin and J. M. Hilbe. *Generalized Linear Models and Extensions*. Stata Press, College Station, 3 edition, 2012.

Daniel F. Heitjan and Donald B. Rubin. Inference from coarse data via multiple imputation with application to age heaping. *Journal of the American Statistical Association*, 85:304–314, 1990.

N. Johnson, S. Kotz, and A.W. Kemp. *Univariate Discrete Distributions*. Wiley, New York, 1992.

R. C. Klesges, M. Debon, and J. W. Ray. Are self-reports of smoking rate biased? evidence from the second national health and nutrition examination survey. *Journal of Clinical Epidemiology*, 48:1225–1233, 1995.

Diana Lambert. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34:1–14, 1992.

Jerald F. Lawless. Negative binomial and mixed poisson regression. *The Canadian Journal of Statistics*, 15:209–225, 1987.

Y. Lee and J. A. Nelder. Two ways of modelling overdispersion in non-normal data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 49:591–598, 2000.

J. M. Lewis-Esquerre, S. M. Colby, T. O. Tevyaw, C. A. Eaton, C. W. Kahler, and P. M. Monti. Validation of the timeline follow-back in the assessment of adolescent smoking. *Drug and Alcohol Dependence*, 79:33–43, 2005.

R. Myers. Errors and bias in the reporting of ages in census data. *Trans Actuarial Soc Am*, 41(2):411–415, 1940.

Paul J. Nietert, Andrea M. Wessell, Chris Feifer, and Steven M. Ornstein. Effect of terminal digit preference on blood pressure measurement and treatment in primary care. *American Journal of Hypertension*, 19:147–152, 2006.

. NIMH Multisite HIV/STD Prevention Trial for African American Couples Group. Methodological overview of an african american couple-based hiv/std prevention trial. *J Acquir Immune Defic Syndr*, 49(Suppl 1):S3–14, 2008.

Geeta S. Pardeshi. Age heaping and accuracy of age data collected during a community survey in the yavatmal district, maharashtra. *Indian Journal of Community Medicine*, 35:391–395, 2010.

Karen L. Price and John W. Seaman. Bayesian modeling of retrospective time-to-pregnancy data with digit preference bias. *Mathematical and Computer Modelling*, 43:1424–1433, 2006.

Pedro Puig and Jordi Valero. Count data distributions. *Journal of the American Statistical Association*, 101:332–340, 2006.

Martin Ridout, Clarice G. B. Demetrio, and John Hinde. Models for count data with many zeros. *Proceedings of the XIXth International Biometrics Conference*, pages 179–192, 1998.

Martin S. Ridout and Byron J. T. Morgan. Modelling digit preference in fecundability studies. *Biometrics*, 47:1423–1433, 1991.

John M. Roberts and Devon D. Brewer. Measures and tests of heaping in discrete quantitative distributions. *Journal of Applied Statistics*, 28:887–896, 2001.

S Thavarajah, WB White, and GA Mansoor. Terminal digit bias in a specialty hypertension faculty practice. *Journal of Human Hypertension*, 17:819–822, 2003.

Shahid Ullah, Caroline F. Finch, and Lesley Day. Statistical modelling for falls count data. *Accident Analysis and Prevention*, 42:384–392, 2010.

Quang Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2):307–333, 1989.

Hao Wang and Daniel F. Heitjan. Modeling heaping in self-reported cigarette counts. *Statistics in Medicine*, 27:3789–3804, 2008.

J. Wolff and T. Augustin. Heaping and its consequences for duration analysis–a simulation study. *Allgemeines Statistisches Archiv*, 87:1–28, 2003.

Z. Yang, J.W. Hardin, and C.L. Addy. A score test for overdispersion in poisson regression based on the generalized poisson-2 model. *Journal of Statistical Planning and Inference*, 139:1514–1521, 2009.

Appendix A

1st derivatives of Heaped Distributions

1st derivatives of Heaped Generalized Poisson Distribution

$$
\begin{aligned}
A_1 &= \Gamma_{\mathrm{R}}(y\alpha - h, \mu) - \Gamma_{\mathrm{R}}(y\alpha + h + 1, \mu) \\
A_2 &= \Gamma_{\mathrm{R}}(y\alpha + h + 1, \mu)\left[\psi(y\alpha + h + 1) - \log \mu\right] \\
&\quad + \Gamma_{\mathrm{R}}(y\alpha - h, \mu)\left[-\psi(y\alpha - h) + \log \mu\right] \\
M_{\mathrm{G1}} &= \mathrm{Meijer}_{\mathrm{G}}\left(\{\{\}, \{1, 1\}\}, \{\{0, 0, y\alpha - h\}, \{\}\}, \mu\right) \\
M_{\mathrm{G2}} &= \mathrm{Meijer}_{\mathrm{G}}\left(\{\{\}, \{1, 1\}\}, \{\{0, 0, y\alpha + h + 1\}, \{\}\}, \mu\right)
\end{aligned}
$$

where Γ_{R} is the gamma regularized function, μ is the link function, ψ is the digamma function, and h is half-width of heaping interval.

$$
\frac{\partial \mathcal{L}}{\partial \beta^T} = \frac{1}{A_1}\left[e^{-\mu}\left(\frac{\mu^{y\alpha+h+1}}{\Gamma(y\alpha + h + 1)} - \frac{\mu^{y\alpha-h}}{\mu\Gamma(y\alpha - h)}\right)\mu\right]
$$

$$
\frac{\partial \mathcal{L}}{\partial \alpha^T} = \frac{y}{A_1}\left[\frac{M_{\mathrm{G1}}}{\Gamma(y\alpha - h)} + \frac{A_2\Gamma(y\alpha + h + 1) - M_{\mathrm{G2}}}{\Gamma(y\alpha + h + 1)}\right]
$$

$$
\begin{aligned}
B_1 &= \text{Beta}\left[1+h-y, \frac{1}{1+\alpha\mu}, \alpha\right] \\
B_2 &= \text{Beta}\left[-h-y, \frac{1}{1+\alpha\mu}, \alpha\right] \\
B_3 &= \Big((1+\alpha\mu)^2 \text{Beta}(\alpha, \frac{1}{1+\alpha\mu})\Big)\Big(\text{Beta}_{\text{R}}\left[-h+y, \alpha, \frac{1}{1+\alpha\mu}\right] \\
&\quad -\text{Beta}_{\text{R}}\left[1+h+y, \alpha, \frac{1}{1+\alpha\mu}\right]\Big) \\
D_1 &= -\psi\left[\frac{1}{1+\alpha\mu}\right] + \psi\left[\alpha + \frac{1}{1+\alpha\mu}\right] + \text{Log}[-h-y] \\
D_2 &= -\psi\left[\frac{1}{1+\alpha\mu}\right] + \psi\left[\alpha + \frac{1}{1+\alpha\mu}\right] + \text{Log}[1+h-y] \\
H_1 &= \text{HPFQReg}\left[\left\{\frac{1}{1+\alpha\mu}, \frac{1}{1+\alpha\mu}, 1-\alpha\right\}, \left\{1+\frac{1}{1+\alpha\mu}, 1+\frac{1}{1+\alpha\mu}\right\}, -h-y\right] \\
H_2 &= \text{HPFQReg}\left[\left\{\frac{1}{1+\alpha\mu}, \frac{1}{1+\alpha\mu}, 1-\alpha\right\}, \left\{1+\frac{1}{1+\alpha\mu}, 1+\frac{1}{1+\alpha\mu}\right\}, 1+h-y\right] \\
H_3 &= \text{HPFQReg}\left[\left\{\alpha, \alpha, \frac{\alpha\mu}{1+\alpha\mu}\right\}, \{1+\alpha, 1+\alpha\}, -h+y\right] \\
H_4 &= \text{HPFQReg}\left[\left\{\alpha, \alpha, \frac{\alpha\mu}{1+\alpha\mu}\right\}, \{1+\alpha, 1+\alpha\}, 1+h+y\right]
\end{aligned}
$$

where Beta_{R} is the beta regularized function, μ is the link function, HPFQReg is the HypergeometricPFQRegularized function, and h is half-width of heaping interval.

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \beta^T} = \frac{\alpha}{B_3}\Bigg[&(-y-h)^{\frac{1}{1+\alpha\mu}}\Gamma\Big(\frac{1}{1+\alpha\mu}\Big)^2 H_1 - (1-y+h)^{\frac{1}{1+\alpha\mu}}\Gamma\Big(\frac{1}{1+\alpha\mu}\Big)^2 H_2 \\
&- B_2\left[D_1\right] + B_1\left[D_2\right]\mu
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \alpha^T} = & \frac{1}{\left(\mathrm{Beta_R}\left[-h+y,\alpha,\frac{1}{1+\alpha\mu}\right] - \mathrm{Beta_R}\left[1+h+y,\alpha,\frac{1}{1+\alpha\mu}\right]\right)} \\
& \Bigg[- \frac{(y-h)^{\alpha}\Gamma(\alpha)^2 H_3 + (1+y+h)^{\alpha}\Gamma(\alpha)^2 H_4}{\mathrm{Beta}\left[\alpha,\frac{1}{1+\alpha\mu}\right]} \\
& + \frac{1}{(1+\alpha\mu)^2 \mathrm{Beta}(\alpha,\frac{1}{1+\alpha\mu})\mu\left[(-y-h)^{\frac{1}{1+\alpha\mu}}\Gamma(\frac{1}{1+\alpha\mu})^2 H_1 - B_2 D_1\right]} \\
& - \frac{1}{1+\alpha\mu)^2 \mathrm{Beta}(\alpha,\frac{1}{1+\alpha\mu})\mu\left[(-y+h+1)^{\frac{1}{1+\alpha\mu}}\Gamma(\frac{1}{1+\alpha\mu})^2 H_2 - B_1 D_2\right]} \\
& + \mathrm{Beta_R}\left[-h+y,\alpha,\frac{1}{1+\alpha\mu}\right]\left(-\psi(\alpha) + \psi\left[\alpha + \frac{1}{1+\alpha\mu}\right] + \mathrm{Log}[y-h]\right) \\
& - \mathrm{Beta_R}\left[1+h+y,\alpha,\frac{1}{1+\alpha\mu}\right]\left(-\psi(\alpha) + \psi\left[\alpha + \frac{1}{1+\alpha\mu}\right] + \mathrm{Log}[1+h-y]\right]
\end{aligned}
$$

Appendix B

1st derivatives of Heaped Zero-Inflated Distributions

1st derivatives of Heaped Zero-Inflated Generalized Poisson Distribution

$$
\begin{aligned}
A_1 &= \Gamma_{\mathrm{R}}(y\alpha - h, \mu) - \Gamma_{\mathrm{R}}(y\alpha + h + 1, \mu) \\
A_2 &= \Gamma_{\mathrm{R}}(y\alpha + h + 1, \mu)\,[\psi(y\alpha + h + 1) - \log\mu] \\
&\quad +\Gamma_{\mathrm{R}}(y\alpha - h, \mu)\,[-\psi(y\alpha - h) + \log\mu] \\
M_{\mathrm{G1}} &= \mathrm{Meijer}_{\mathrm{G}}\left(\{\{\}, \{1,1\}\}, \{\{0,0,y\alpha - h\}, \{\}\}, \mu\right) \\
M_{\mathrm{G2}} &= \mathrm{Meijer}_{\mathrm{G}}\left(\{\{\}, \{1,1\}\}, \{\{0,0,y\alpha + h + 1\}, \{\}\}, \mu\right)
\end{aligned}
$$

where Γ_{R} is the gamma regularized function, μ is the link function, ψ is the digamma function, w is the binary distribution for the probability of a zero outcome, and h is half-width of heaping interval.

$$
\frac{\partial \mathcal{L}}{\partial \beta^T} = (-\exp(-1+w)\mu) + \frac{1}{A_1}\left[e^{-\mu}(1-w)\left(\frac{\mu^{y\alpha+h+1}}{\Gamma(y\alpha+h+1)} - \frac{\mu^{y\alpha-h}}{\mu\Gamma(y\alpha-h)}\right)\mu\right]
$$

$$
\frac{\partial \mathcal{L}}{\partial \alpha^T} = \frac{(1-w)y}{A_1}\left[\frac{M_{\mathrm{G1}}}{\Gamma(y\alpha-h)} + \frac{A_2\Gamma(y\alpha+h+1) - M_{\mathrm{G2}}}{\Gamma(y\alpha+h+1)}\right]
$$

1ST DERIVATIVES OF HEAPED ZERO-INFLATED NEGATIVE BINOMIAL DISTRIBUTION

$$
\begin{aligned}
B_1 &= \text{Beta}\left[1+h-y, \frac{1}{1+\alpha\mu}, \alpha\right] \\
B_2 &= \text{Beta}\left[-h-y, \frac{1}{1+\alpha\mu}, \alpha\right] \\
B_3 &= \left((1+\alpha\mu)^2 \text{Beta}(\alpha, \frac{1}{1+\alpha\mu})\right)\left(\text{Beta}_{\text{R}}\left[-h+y, \alpha, \frac{1}{1+\alpha\mu}\right]\right. \\
&\quad \left.-\text{Beta}_{\text{R}}\left[1+h+y, \alpha, \frac{1}{1+\alpha\mu}\right]\right) \\
D_1 &= -\psi\left[\frac{1}{1+\alpha\mu}\right] + \psi\left[\alpha + \frac{1}{1+\alpha\mu}\right] + \text{Log}[-h-y] \\
D_2 &= -\psi\left[\frac{1}{1+\alpha\mu}\right] + \psi\left[\alpha + \frac{1}{1+\alpha\mu}\right] + \text{Log}[1+h-y] \\
H_1 &= \text{HPFQReg}\left[\left\{\frac{1}{1+\alpha\mu}, \frac{1}{1+\alpha\mu}, 1-\alpha\right\}, \left\{1+\frac{1}{1+\alpha\mu}, 1+\frac{1}{1+\alpha\mu}\right\}, -h-y\right] \\
H_2 &= \text{HPFQReg}\left[\left\{\frac{1}{1+\alpha\mu}, \frac{1}{1+\alpha\mu}, 1-\alpha\right\}, \left\{1+\frac{1}{1+\alpha\mu}, 1+\frac{1}{1+\alpha\mu}\right\}, 1+h-y\right] \\
H_3 &= \text{HPFQReg}\left[\left\{\alpha, \alpha, \frac{\alpha\mu}{1+\alpha\mu}\right\}, \{1+\alpha, 1+\alpha\}, -h+y\right] \\
H_4 &= \text{HPFQReg}\left[\left\{\alpha, \alpha, \frac{\alpha\mu}{1+\alpha\mu}\right\}, \{1+\alpha, 1+\alpha\}, 1+h+y\right]
\end{aligned}
$$

where Beta_{R} is the beta regularized function, μ is the link function, HPFQReg is the HypergeometricPFQRegularized function, and h is half-width of heaping interval.

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \beta^T} = &((-1+w)\alpha(-1+\mu)(1+\alpha\mu)^{-2-\frac{1}{\alpha}}\mu) \\
&+ \frac{\alpha}{B_3}\left[(-y-h)^{\frac{1}{1+\alpha\mu}}\Gamma\left(\frac{1}{1+\alpha\mu}\right)^2 H_1 - (1-y+h)^{\frac{1}{1+\alpha\mu}}\Gamma\left(\frac{1}{1+\alpha\mu}\right)^2 H_2\right. \\
&- B_2\left[D_1\right] + B_1\left[D_2\right]\mu
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \alpha^{T}} = & -\frac{(-1+w)\mu(1+\alpha\mu)^{-2-\frac{1}{\alpha}}(\alpha-\alpha\mu+(1+\alpha\mu)\mathrm{Log}[1+\alpha\mu])}{\alpha} \\
& + \frac{(1-w)}{\Big(\mathrm{Beta}_{\mathrm{R}}\big[-h+y,\alpha,\tfrac{1}{1+\alpha\mu}\big]-\mathrm{Beta}_{\mathrm{R}}\big[1+h+y,\alpha,\tfrac{1}{1+\alpha\mu}\big]\Big)} \\
& \Big[-\frac{(y-h)^{\alpha}\Gamma(\alpha)^{2}H_{3}+(1+y+h)^{\alpha}\Gamma(\alpha)^{2}H_{4}}{\mathrm{Beta}\big[\alpha,\tfrac{1}{1+\alpha\mu}\big]} \\
& + \frac{1}{(1+\alpha\mu)^{2}\mathrm{Beta}(\alpha,\tfrac{1}{1+\alpha\mu})\mu\big[(-y-h)^{\frac{1}{1+\alpha\mu}}\Gamma(\tfrac{1}{1+\alpha\mu})^{2}H_{1}-B_{2}D_{1}\big]} \\
& - \frac{1}{1+\alpha\mu)^{2}\mathrm{Beta}(\alpha,\tfrac{1}{1+\alpha\mu})\mu\big[(-y+h+1)^{\frac{1}{1+\alpha\mu}}\Gamma(\tfrac{1}{1+\alpha\mu})^{2}H_{2}-B_{1}D_{2}\big]} \\
& + \mathrm{Beta}_{\mathrm{R}}\left[-h+y,\alpha,\frac{1}{1+\alpha\mu}\right]\left(-\psi(\alpha)+\psi\left[\alpha+\frac{1}{1+\alpha\mu}\right]+\mathrm{Log}[y-h]\right) \\
& - \mathrm{Beta}_{\mathrm{R}}\left[1+h+y,\alpha,\frac{1}{1+\alpha\mu}\right]\left(-\psi(\alpha)+\psi\left[\alpha+\frac{1}{1+\alpha\mu}\right]+\mathrm{Log}[1+h-y]\right]
\end{aligned}
$$