2018

Adjusting For Mis-Reporting In Count Data

Gelareh Rahimighazikalayeh
University of South Carolina - Columbia

ADJUSTING FOR MIS-REPORTING IN COUNT DATA

by

Gelareh Rahimighazikalayeh

Bachelor of Science
Shahid Beheshti University, 2008

Master of Science
Shahid Beheshti University of Medical Sciences, 2011

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

2018

Accepted by:

James W. Hardin, Major Professor

James Hussey, Committee member

Feifei Xiao, Committee member

Mindi Spencer, Committee member

Cheryl L. Addy, Vice Provost and Dean of the Graduate School

DEDICATION

I would like to dedicate my dissertation work to my loving parents, Jabbareh and Manouchehr whose word of encouragement and push for tenacity ring in my ears. A special feeling of gratitude to my sisters Delaram and Azadeh who have been a constant source of support and encouragement during the challenges of graduate school and life.

ACKNOWLEDGMENTS

I would like to express the deepest appreciation to my dissertation advisor, Dr. James W. Hardin for the continuous support of my PhD study, for his patience and immense knowledge. As my teacher and mentor, he has taught me more than I could give him credit for here.

I would like to thank my committee members, Drs. James Hussey, Feifei Xiao and Mindi Spencer for their continued support and for agreeing to serve on my dissertation committee on short notice.

I am grateful to all of those with whom I have had the pleasure to work with during my PhD program. I would especially like to thank Dr. Maggie Miller who has provided me extensive personal and professional guidance and taught me a great deal about both scientific research and life in general.

ABSTRACT

Any counting system is prone to recording errors including underreporting and overreporting. Ignoring the misreporting pattern in count data can give rise to bias in the estimation of model parameters. Accordingly, Poisson, negative binomial and generalized Poisson regression have been expanded in some instances to capture reporting biases. However, to our knowledge, no program has been developed to allow users to apply all of these models when needed. In the first part of the dissertation, we review the available models for underreported counts and develop a Stata command to estimate Poisson, negative binomial and generalized Poisson regression models for underreported data.

Although considerable research has been devoted to underreporting models, less attention has been given to inflated counts. Based on the structural model proposed by Li et al. (2003), we will develop two models applicable to potentially misreported data. The first model covers situations where both the reported counts and the true counts follow a Poisson distribution. The second model would be relevant to cases where the actual-unobserved counts are assumed to be from a generalized Poisson distribution and the reported counts are from a Poisson distribution.

The proposed models adjust for both overreporting and underreporting. Our approach allows users to specify the individual's characteristics that contribute to misreporting. With only observed counts at hand, our proposed models estimate the proportions of under/overreporting conditionally.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1
INTRODUCTION

There are many contexts in which the outcome of interest or the dependent variable is a count, for example, the number of occurrences of an event. Inherently, a count variable only takes non-negative integer values. As a result, the distribution of the outcome is usually positively skewed (especially when the mean is small). Similar to classic regression, in count data analysis we wish to explain the outcome of interest through a set of covariates. However, since one of the main assumptions of linear models is heteroscedasticity of the error, standard regression models cannot be applied to count data.

Regression models for counts, like other limited or discrete dependent variable models such as the logit and probit, are non-linear with many properties and special features intimately connected to discreteness and non-linearity (Cameron & Trivedi, 2001). Some of these regression models have been applied to data on number of live births over a specified age interval of the mother (Winkelmann, 1995), number of accidents experienced by an airline over some period (Rose, 1990) or number of times that individuals utilize a health service, such as number of visits to a doctor in the past year (Cameron, Trivedi, Milne, & Piggott, 1988). In most of these cases, the number of counts could have been potentially overreported, underreported or correctly reported. In the case of the counts having been correctly reported, the appropriate count data regression model such as negative binomial, Poisson and generalized Poisson can be

applied on such data. In real life there is potential of misreporting and it is necessary to check count data for this kind of reporting (Pararai, Famoye, & Lee, 2010).

Underreporting is a problem in data collection that occurs when the counting of some event is for some reason incomplete. Any reporting or counting system is prone to such errors in recording. The reasons may be quite different in the various fields of application like public health, criminology, actuarial science or production. In public health we have reporting systems for infectious diseases like HIV or chronic diseases like diabetes in which recording failures may occur as result of diagnostic errors or patients avoiding diagnosis. The same holds for traffic accidents with minor damage. Insurance companies are faced with an unknown number of total claims, as some claims are made with a delay that may be as long as five years. An example from industrial production is the number of products that are broken within a certain period, typically the warranty period. To know this number is important for quality management. Only the number of returned products is known, but the true total number includes also those goods that are not returned by customers. In all these cases reporting systems give lower counts than the actual number of events. Therefore, underreporting is a widespread phenomenon and the estimation of the total number of cases is of particular interest (Neubauer, Duras, & Friedl, 2010).

Overreporting in registration systems occurs when the reported number of events is higher than the actual counts. Depending on the field of application, various factors might play a role in overreporting of an event. In public health, a physicians' mistakes in the diagnostic process could result in over reporting of a specific disease. An example from survey research could be overreporting hand washing behavior in hospital settings

(Contzen, De Pasquale, & Mosler, 2015). Two different explanations of overreporting have been tendered with regard to survey responses. One explanation considers inaccurate memory function or recall errors and the second is social desirability which has been claimed to be the main cause of inflated self-reports (Contzen et al., 2015). In general, research participants want to respond in a way that makes them look as good as possible. Thus, they tend to under-report behaviors deemed inappropriate by researchers or other observers, and they tend to over-report behaviors viewed as appropriate (Donaldson & Grant-Vallone, 2002).

Several methods have been proposed by various authors to address the misreporting problem in count data (Fader & Hardie, 2000; Mukhopadhyay & Trivedi, 1997; Neubauer & Djuraš, 2008, 2009; Winkelmann, 1996). While most of the available methods focus on adjusting for underreported counts, there exist a couple of models that also incorporate overreported data.

In this dissertation, I will review the available models for underreported counts in Chapter 2 and present Stata estimation commands for each, followed by a simulation study to show the performance of the program. In Chapter 3, two models for misreported counts will be introduced, a Poisson mixture model and a generalized Poisson mixture model which adjust for both underreporting and overreporting. The performance of the proposed models will be examined through a simulation studies. A real data analysis will be carried out in Chapter 4 using EBAN study data, An HIV/STD Intervention for African American Couples. I conclude the dissertation in Chapter 5 with future research ideas and applications.

CHAPTER 2
REGRESSION MODELS FOR UNDERREPORTED COUNTS

Underreported count data are generated when only a fraction of the actual events of interest are reported. Let y_i^* denote the total number of events during a fixed time period t for individual i. Suppose that y_i , the observed counts, conditional on y_i^* is characterized by a conditional binomial distribution given by

$$P(y_i|y_i^*, p_i) = \frac{y_i^*!}{(y_i^* - y_i)!\ y_i!}\ p_i^{y_i}\ (1-p_i)^{y_i^* - y_i} \tag{2.1}$$

where p_i gives the individual probability of reporting an event. This probability is assumed to be constant and identical for all events and independent of the history of the process. A given number of the reported events can then arise in many ways. For instance, if $y_i = y_i^*$ then all the events are accurately reported. Alternatively, $y_i = y_i^* - c$ where $0 < c < y_i^*$ can be any number of non-reported events. Most of the models for underreported count data work within this basic framework.

2.1 POISSON MODEL FOR UNDERREPORTING

Winkelmann (Winkelmann, 1996) proposed a mixture of the Poisson and the binomial distributions to take underreporting into account. In this misture model, the true number of events, y_i^* , is assumed to have a Poisson distribution with conditional mean parameterized as

$$E(y_i^*|\boldsymbol{x}_i) = \mu_i = \exp(\boldsymbol{x}_i\boldsymbol{\beta}) \quad , \quad i = 1, \dots, n \tag{2.2}$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)'$ is a vector of unknown regression coefficients and $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ik})$ includes covariates of interest. Assuming a binomial distribution for the observed counts, conditional on y_i^* (2.1), the marginal distribution of the number of reported events y_i can be calculated as

$$P(y_i = y) = \sum_{y^* \geq y}^{\infty} \frac{\mu^{y^*} e^{-\mu}}{y^*!} \frac{y^*!}{(y^* - y)!} p^y (1-p)^{y^* - y}$$

$$= \frac{e^{-\mu p} (\mu p)^y}{y!} \tag{2.3}$$

Hence, the number of observed events is again Poisson distributed with mean $\lambda_i = \mu_i p_i$.

According to Winkelmann (Winkelmann, 1996), if we can capture the structure of the relationship between the observed counts and the actual counts, i.e. the cross-sectional heterogeneity, then the parameters μ and p are both identifiable. Once the model is specified, it is often possible to make conditional statements about each individual's unobserved but true number of events based on their reported counts. There are three conditional distributions that may be of interest:

- First is $P(y^* = a|y = b)$, i.e. the probability of someone having been involved in a events, conditional on the fact that they reported b such of events.
- Second is $g(p|y = b)$, i.e. the distribution of one's reporting probability given that they reported b events.

- The third is $f(\mu|y=b)$, i.e. the distribution of one's true rate parameter, conditional on reporting b events.

Neubauer and Djuraš (Neubauer & Djuraš, 2008, 2009) extended the binomial model for undercounts to the case where both parameters of the binomial model are treated as random. They suggested using mixed models for undercounts to allow for larger variability in the response, i.e. allowing for more overdispersion.

Winkelmann (Winkelmann, 1996) also considered a hierarchical Bayesian approach where the actual counts are modeled through a Poisson regression with a multivariate normal prior on the covariate coefficients and a uniform prior is placed on the reporting probability p. "The problem with this approach is that it is intractable to analytically derive the marginal posterior distribution for the parameters of interest and so computationally intensive Markov Chain Monte Carlo (MCMC) methods were required to make the inference of interest" (Fader & Hardie, 2000).

As an alternative, I used a maximum likelihood method for the estimation process. According to (2.3), the number of observed counts is Poisson distributed with mean $\lambda_i = \mu_i p_i$ so a realistic model is then given by

$$\mu_i = \exp(\acute{\boldsymbol{x}}_i\,\boldsymbol{\beta}) \quad and \quad p_i = \frac{\exp(\acute{\boldsymbol{z}}_i\boldsymbol{\gamma})}{1+\exp(\acute{\boldsymbol{z}}_i\boldsymbol{\gamma})}$$

where $\boldsymbol{x}_i$and $\boldsymbol{z}_i$ are two sets of covariates defining the marginal means, μ_i, and the reporting probability, p_i, respectively. $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the unknown parameters to be estimated The likelihood contribution of the i-th observation is given by

$$L(\beta,\gamma|y_i,x_i,z_i) = \frac{e^{-(\mu_i(\beta)\,p_i(\gamma))}\left(\mu_i(\beta)\,p_i(\gamma)\right)^{y_i}}{y_i!} \qquad (2.4)$$

If we imagine a model in which both $\boldsymbol{x}$ and $\boldsymbol{z}$ consist of a constant term, then there is an infinite number of solutions for which the mean of the true counts is equal to

$$\exp(\beta_0)\frac{\exp(\gamma_0)}{1+\exp(\gamma_0)}$$

Therefore, for identifiability, $\mathbf{z}$ cannot contain a constant term. If we look further at a single common binary covariate, it is just as easy to see there is no identifiable solution. Thus, the covariates in $\boldsymbol{z}$ and $\boldsymbol{x}$ cannot overlap.

We developed a Stata command named "undct" for performing underreporting count data regression. The general syntax of the program is

undct *depvar* [*indepvars*] [*if*] [*in*] [weight],

under (*varlist* [, offset (*varname*)] | _cons) [*options*]

where the distribution of the dependent variable can be specified in the [*options*]. A Poisson-binomial model can be developed by choosing a Poisson distribution for the outcome of interest. In the upcoming section, I will illustrate the undct command for fitting an underreported count regression model to simulated data.

2.2 SIMULATION STUDY FOR POISSON UNDERCOUNT MODEL

We conducted a simulation study to examine the performance of the Poisson-binomial mixture model compared with that of the standard Poisson model. For the parameters to be identifiable, we used two sets of disjoint variables for $\boldsymbol{x}$ and $\boldsymbol{z}$. In every iteration, a data set of size 1000 was synthesized, first a Poisson model with covariates in $\boldsymbol{x}$ was fit and then a Poisson-binomial mixture model was applied to the synthesized data. These procedures were repeated 100 times independently for each of the following scenarios.

- x_1 follows a standard normal, x_2 is a binary variable with $p = 0.5$, z_1 is a random uniform variable on $(0,1)$ and z_2 follows a binary distribution with $p = 0.3$.

  $\mu_i = \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})$

  $\text{logit}(p_i) = \gamma_1 z_{1i} + \gamma_2 z_{2i}$

  we used the following parameter values:

  $\beta_0 = 1.5, \beta_1 = -0.5, \beta_2 = -1, \gamma_1 = 0.3, \gamma_2 = 0.7$

- x_1 follows a standard normal, x_2 is a binary variable with $p = 0.5$, x_3 is Poisson distributed with mean 2, z_1 is a random uniform variable on $(0,1)$, z_2 follows a binary distribution with $p = 0.3$.

  $\mu_i = \exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i})$

  $\text{logit}(p_i) = \gamma_1 z_{1i} + \gamma_2 z_{2i}$

  we used the following parameter values:

  $\beta_1 = 0.4,\ \beta_2 = 0.8, \beta_3 = -0.3, \gamma_1 = -0.5, \gamma_2 = 1.5$

The results are summarized in Tables 2-1 and 2-2.

Table 2.1 Simulation results for Poisson model

First simulation scenario				Second simulation scenario			
True value	Mean	SD	Bias	True Value	Mean	SD	Bias
$\beta_0 = 1.5$	0.961	0.030	0.538	$\beta_1 = 0.4$	0.343	0.039	0.056
$\beta_1 = -0.5$	-0.502	0.021	0.002	$\beta_2 = 0.8$	0.394	0.054	0.405
$\beta_2 = -1$	-0.992	0.052	-0.007	$\beta_3 = -0.3$	-0.440	0.027	0.140

In the first simulation scenario, the logarithm of the marginal means was explained through a constant and two regressors, x_1 and x_2. The reporting probability of each event

Table 2.2 Simulation results for Poisson undercount model

First simulation scenario				Second simulation scenario			
True value	Mean	SD	Bias	True Value	Mean	SD	Bias
$\beta_0 = 1.5$	1.502	0.057	-0.002	$\beta_1 = 0.4$	0.396	0.044	0.003
$\beta_1 = -0.5$	-0.501	0.021	0.001	$\beta_2 = 0.8$	0.794	0.070	0.005
$\beta_2 = -1$	-0.991	0.054	-0.008	$\beta_3 = -0.3$	-0.299	0.030	0.000
$\gamma_1 = 0.3$	0.273	0.195	0.026	$\gamma_1 = -0.5$	-0.479	0.216	-0.020
$\gamma_2 = 0.7$	0.708	0.177	-0.008	$\gamma_2 = 1.5$	1.545	0.433	-0.045

was also assumed to be related to the explanatory variables z_1 and z_2 through a logit link function. Both, the classic Poisson regression and the Poisson-binomial mixture model provided good estimates of the effects of x_1 and x_2 on the observed counts. However, this was not the same for the intercept. The standard Poisson model estimated the baseline incidence rate ratio (IRR) to be $\exp(0.961) = 2.614$, while the actual value was $\exp(1.5) = 4.481$. In contrast, the undercount model was able to precisely capture the effects of all covariates. This suggests that, when underreporting is present, the Poisson regression is likely to be misleading due to biased results it provides for the model's intercept.

In the second simulation scenario, we were interested to compare the two discussed regression approaches when the constant is excluded from the models. So, we related the true means to three regressors and we considered two covariates for explaining the reporting probability. Not surprisingly, all the estimated coefficients from the Poisson model were biased. The IRRs produced by this model were 1.409, 1.482 and

0.644 for x_1, x_2 and x_3 respectively when the actual values were 1.491, 2.225 and 0.740. On the other hand, the undercount model estimated the IRRs as 1.485, 2.212 and 0.741.

On the basis of the simulation results reported in Tables 2.1 and 2.2, we conclude that the conventional Poisson regression can suffer from model misspecification when used to model underreported data. The Poisson-binomial mixture model, on the other hand, can provide reliable estimates in this context. They can also provide information on the association of potential covariates with reporting probability of the events.

2.3 NEGATIVE BINOMIAL MODEL FOR UNDERREPORTING

One of the most important features about Poisson regression is the equidispersion assumption. In research, however, collected count data often displays heterogeneity across observational units that exceed the assumed conditional variance. It can be shown that wrongly assuming equidispersion might affect the robustness of estimators produced by Poisson model which consequently leads to misleading inferences about the regression. Among models that have been introduced to overcome this problem, the negative binomial regression is the most commonly used alternative to the Poisson regression when overdispersion is present.

According to Winkelman (Winkelmann, 1996), the Poisson-binomial model for underreporting and the Poisson model with unobserved heterogeneity share similar structural properties in the sense that random underreporting also leads to overdispersion in the observed counts. However, it is hard to disentangle overdispersion due to underreporting from that of unobserved heterogeneity. A negative binomial regression that can further capture underreporting can be a natural remedy to attack this problem.

In 1997, Mukhopadhay (Mukhopadhyay & Trivedi, 1997) extended the model proposed by Winkelmann (Winkelmann, 1996) to situations where the true counts follow a negative binomial Distribution. In the underreporting context, the construction of the negative binomial model can be made based on the following assumptions:

i. For each individual, the actual number of events, y_i^*, in a unit time interval is Poisson distributed with mean μ_i

$$P(Y^* = y_i^* | \mu_i) = \frac{e^{-\mu_i}\, \mu_i^{y_i^*}}{y_i^*!} \qquad , i = 1, \dots, n \tag{2.5}$$

ii. The distribution of μ_i across individuals is gamma with parameters (θ, θ)

$$f(\mu_i) = \frac{\theta^\theta}{\Gamma(\theta)} \mu_i^{\theta-1} e^{-\theta\mu_i} \qquad , \theta > 0 \tag{2.6}$$

iii. Conditional on y_i^*, the observed counts have a binomial distribution with parameters (y_i^*, p_i)

$$P(Y = y_i | y_i^*, p_i) = \binom{y_i^*}{y_i} p_i^{y_i} (1 - p_i)^{y_i^* - y_i} \tag{2.7}$$

iv. An individual's reporting probability, p_i, is independent of their marginal mean, μ_i .

Combining the assumptions (i) and (ii) gives us the marginal distribution of the actual counts which is a negative binomial with mean μ_i and dispersion parameter α, where $\alpha = {}^1/_\theta$

$$P(Y^* = y_i^*) = \int_0^\infty P(Y^* = y_i^* | \mu_i)\, f(\mu_i)\, d\mu_i =$$

$$\frac{\Gamma(y_i^* + \alpha^{-1})}{y_i^*!\ \Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i^*} \tag{2.8}$$

In a similar manner, combining the result (2.8) with assumption (iii) give us the marginal distribution of the observed counts. Mukhopadhay (Mukhopadhyay & Trivedi, 1997) has derived this distribution as

$$P(Y = y_i) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i + 1)\ \Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+p_i\mu_i}\right)^{\alpha^{-1}}\left(\frac{p_i\mu_i}{\alpha^{-1}+p_i\mu_i}\right)^{y_i} \tag{2.9}$$

Thus, the marginal distribution of the observed counts is again negative binomial with mean and variance equal to $p_i\mu_i$ and $p_i\mu_i(1 + \alpha p_i\mu_i)$.

Similar to the Poisson-binomial mixture model discussed in section 2.1, we can model μ_i and p_i through some explanatory variables. Let $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ik})$ be a set of covariates defining the marginal means, μ_i, and $\boldsymbol{z}_i = (z_{i1}, \dots, z_{il})$ be a separate set of regressors affecting the reporting probability, p_i, then

$$\mu_i = \exp(\boldsymbol{x}_i\,\boldsymbol{\beta}) \quad and \quad p_i = \frac{\exp(\boldsymbol{z}_i\boldsymbol{\gamma})}{1 + \exp(\boldsymbol{z}_i\boldsymbol{\gamma})}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the unknown parameters to be estimated. Maximum likelihood methods can be used for the estimation purposes.

I use the undct command introduced in Section 2.1 for estimating the negative binomial undercount model. The general syntax of the program is given by

undct *depvar* [*indepvars*] [*if*] [*in*] [weight],

under (*varlist* [, offset (*varname*)] | _cons) [*options*]

where the distribution of the dependent variable should be specified as negative binomial in the [*options*].

2.4 SIMULATION STUDY FOR NEGATIVE BINOMIAL MODEL FOR UNDERREPORTED COUNTS

We conducted a simulation study to examine the performance of the negative binomial undercount model compared with that of the standard negative binomial regression. For the parameters to be identifiable, we used two non-overlapping set of variables for $\boldsymbol{x}$ and $\boldsymbol{z}$. In every iteration, a data set of size 10,000 was synthesized. First a negative binomial model with covariates in $\boldsymbol{x}$ was fit, and then a negative binomial undercount model was applied to the synthesized data. These procedures were repeated 100 times independently for each of the following scenarios:

- x_1 follows a standard normal, x_2 is a binary variable with $p = 0.8$, z_1 is a random uniform variable on $(0,1)$ and z_2 follows a binary distribution with $p = 0.4$. We chose the dispersion parameter, α, to be equal to 0.3.

  $\mu_i = \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})$

  $\text{logit}(p_i) = \gamma_1 z_{1i} + \gamma_2 z_{2i}$

  we have used the following parameter values:

  $\beta_0 = 1.3, \beta_1 = -0.4, \beta_2 = -0.7, \gamma_1 = 0.5, \gamma_2 = 0.9$

- x_1 follows a standard normal, x_2 is a binary variable with $p = 0.8$, x_3 is Poisson distributed with mean 3, z_1 is a random uniform variable on $(0,1)$, z_2 follows a binary distribution with $p = 0.4$. We chose the dispersion parameter, α, to be equal to 0.3.

  $\mu_i = \exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i})$

  $\text{logit}(p_i) = \gamma_1 z_{1i} + \gamma_2 z_{2i}$

  we have used the following parameter values:

$\beta_0 = -0.6,\ \beta_1 = 1.1, \beta_2 = 0.3, \gamma_1 = -0.5, \gamma_2 = 1.5$

The results are summarized in Tables 2-3 and 2-4. In the first simulation scenario, the exponential function was used as a link between the marginal means and the covariates x_1 and x_2. The reporting probability of each event was also regressed on explanatory predictors z_1 and z_2 through a logit link function. Standard negative binomial model provided accurate estimates of both the dispersion parameter and the effects of x_1 and x_2 on the observed counts. But, the estimated value for the intercept was biased. Based on the NB model we predicted the baseline incidence rate to be exp(0.855)=2.351 while the actual value was 3.669. On the other hand, the negative binomial undercount model was able to accurately estimate all coefficients and further provide insight into revealing the underlying factors that contribute to underreporting. This suggests that, when underreporting is present, the negative binomial regression might lead to misleading inferences due to biased estimates it produces for the model's intercept.

Table 2.3 Simulation results for negative binomial model

First simulation scenario				Second simulation scenario			
True value	Mean	SD	Bias	True Value	Mean	SD	Bias
$\beta_0 = 1.3$	0.855	0.021	0.445	$\beta_1 = -0.6$	-0.579	0.008	-0.021
$\beta_1 = -0.4$	-0.399	0.009	-0.001	$\beta_2 = 1.1$	0.723	0.014	0.377
$\beta_2 = -0.7$	-0.701	0.024	0.001	$\beta_3 = 0.3$	0.246	0.003	0.054
$\alpha = 0.3$	0.327	0.014	-0.027	$\alpha = 0.3$	0.414	0.009	-0.114

In the second simulation scenario, we were interested to compare the performance of standard and undercount negative binomial models in the absence of an intercept. To do

Table 2.4 Simulation results for negative binomial undercount model

First simulation scenario				Second simulation scenario			
True value	Mean	SD	Bias	True Value	Mean	SD	Bias
$\beta_0 = 1.3$	1.305	0.037	-0.005	$\beta_1 = -0.6$	-0.600	0.007	0.000
$\beta_1 = -0.4$	-0.400	0.009	0.000	$\beta_2 = 1.1$	1.100	0.014	0.000
$\beta_2 = -0.7$	-0.701	0.023	0.001	$\beta_3 = 0.3$	0.300	0.003	0.000
$\gamma_1 = 0.5$	0.485	0.122	0.015	$\gamma_1 = -0.5$	-0.503	0.048	0.003
$\gamma_2 = 0.9$	0.890	0.096	0.010	$\gamma_2 = 1.5$	1.492	0.074	0.008
$\alpha = 0.3$	0.298	0.013	0.002	$\alpha = 0.3$	0.299	0.008	0.001

so, we related the true means to three regressors and we considered two covariates for explaining the reporting probability. Looking at Table 2.3, the results from the negative binomial model are not satisfactory. While the estimated coefficients for x_1 and x_3 are close to their actual values, the estimates for β_2 and α are both biased. Exponentiating the coefficients, we can better see the amount of bias in incidence rate which is a standard tool for interpreting the results in count regression. The Incidence rate ratios corresponding to x_1, x_2 and x_3 were estimated as 0.560, 2.060 and 1.278 when the actual values were 0.548, 3.004 and 1.349. One might argue that the NB model still seems to be fine making inferences about incidence rate ratios considering the fact that not all the estimates were biased. The problem is that in practice it is unclear which effects are going to be affected by model inaccuracy. Predicting future outcomes would also be fallacious since all estimators, regardless of being biased or not, would have their own share on the calculation process and thus the final result would be altered.

Unlike the standard negative binomial model, the undercount model was very efficient in estimating both dispersion and regression parameters.

In conclusion, the standard negative binomial regression can suffer from model misspecification when used to model underreported data. However, negative binomial undercount models can provide reliable estimates in this context. They can further provide information on the association of potential covariates with reporting probability of the events.

2.5 GENERALIZED POISSON MODEL FOR UNDERREPORTING

While Poisson regression is the most convenient method for modeling count data, it is often too restrictive to hold on to the assumption that the variance is equal to the mean. Frequently, data exhibits an overdispersion pattern, with the variance greater than the mean (Ridout & Besbeas, 2004). As we discussed in Section 2.3, negative binomial regression can be used as an alternative to Poisson regression when overdispersion is present.

At the same time, it is recognized that sometimes the variance of the response variable is less than mean. This phenomenon has been referred to as underdispersion in the literature. Weighted Poisson distributions have been applied by several authors to form models that can handle underdispersed count data (Cameron & Johansson, 1997; Del Castillo & Pérez-Casany, 1998; Ridout & Besbeas, 2004). Some alternative approaches aimed at developing models that accommodate both over- and underdispersion have been introduced (Consul & Famoye, 1992; Shmueli, Minka, Kadane, Borle, & Boatwright, 2005). Among these, the generalized Poisson regression

model has obtained more attention due to its flexibility and convenient properties (Famoye, Wulu, & Singh, 2004; Özmen, 2000; Wang & Famoye, 1997). A number of extensions to generalized Poisson regression have merged in recent years (Bae, Famoye, Wulu, Bartolucci, & Singh, 2005; Czado, Erhardt, Min, & Wagner, 2007; Famoye & Wang, 2004). In 2006, Pararai et al. modified the generalized Poisson regression and developed a model to capture underreporting when the outcome follows generalized Poisson distribution (Pararai, Famoye, & Lee, 2006).

The following assumptions are used for building the generalized Poisson regression model for underreported counts (GPRU)

i. For each individual, the actual number of events, y_i^*, in a unit time interval has generalized Poisson distribution (GP) with probability function

$$f(y_i^*, \mu_i, \alpha) = \frac{\mu_i}{1+\alpha\mu_i}\left[\frac{\mu_i(1+\alpha y_i^*)}{1+\alpha\mu_i}\right]^{y_i^*-1} \exp\left[\frac{-\mu_i(1+\alpha y_i^*)}{1+\alpha\mu_i}\right]\frac{1}{y_i^*!} \quad , y_i^* = 0,1,2,\ldots \tag{2.10}$$

where α and μ_i represent, respectively, the dispersion parameter and the mean. The variance of GP model can be calculated through $\mu_i(1+\alpha\mu_i)^2$. The Poisson distribution is a special case of generalized Poisson distribution and the function in (2.10) reduces to Poisson probability function when $\alpha = 0$ (Consul & Famoye, 1992).

ii. Conditional on y_i^*, the observed counts have a binomial distribution with parameters (y_i^*, p_i)

$$P(Y = y_i | y_i^*, p_i) = \binom{y_i^*}{y_i} p_i^{y_i} (1 - p_i)^{y_i^* - y_i} \tag{2.11}$$

iii. An individual's reporting probability, p_i, is independent of his/her marginal mean, μ_i

The marginal distribution of the observed counts can be calculated by combining the assumption (i) and (ii). Pararai et al. (Pararai et al., 2006), derived the generalized Poisson distribution for underreported counts (GPDU) as

$$P(Y = y_i) = \frac{\mu_i(1 - p_i)}{1 + \alpha\mu_i} \times \left[\frac{\mu_i(1 - p_i + \alpha y_i)}{1 + \alpha\mu_i}\right]^{y_i - 1} \exp\left[\frac{-\mu_i(1 - p_i + \alpha y_i)}{1 + \alpha\mu_i}\right] \frac{1}{y_i!} \quad , \quad y_i = 0,1,2, \ldots \tag{2.12}$$

The mean and variance of GPDU can be obtained by

$$E(Y) = E[E(Y|Y^*)] = \mu(1 - p) \tag{2.13}$$

$$Var(Y) = V[E(Y|Y^*)] + E[V(Y|Y^*)] = \mu(1 - p)[(1 + \alpha\mu)^2 + p\mu] \tag{2.14}$$

The marginal means of the true counts, μ_i, and the reporting probability p_i can be both estimated through some explanatory variables. Let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik})$ and $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{il})$ be two disjoint sets of covariates, then μ_i and p_i can be modeled through

$$\mu_i = \exp(\boldsymbol{x}_i \boldsymbol{\beta}) \quad and \quad p_i = \frac{\exp(\boldsymbol{z}_i \boldsymbol{\gamma})}{1 + \exp(\boldsymbol{z}_i \boldsymbol{\gamma})}$$

Where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the unknown parameters to be estimated.

Maximum likelihood methods can be used for estimating parameters of generalized Poisson regression model for underreported counts (GPRU).

I will use the undct command introduced in Section 2.1 for fitting the GPRU model. The general syntax of the program would be

undct *depvar* [*indepvars*] [*if*] [*in*] [weight],

under (*varlist* [, offset (*varname*)] | _cons) [*options*]

where the distribution of the dependent variable should be specified as generalized Poisson in the [*options*].

2.6 SIMULATION STUDY FOR GENERALIZED POISSON REGRESSION MODEL FOR UNDERREPORTED COUNTS

We conducted a simulation study to examine the performance of the GPRU model compared with that of the standard generalized Poisson regression. For the parameters to be identifiable, we used two non-overlapping set of variables for $\boldsymbol{x}$ and $\boldsymbol{z}$. In every iteration, a data set of size 10,000 was synthesized, first a GPR model with covariates in $\boldsymbol{x}$ were fitted and then a GPRU model were applied to the synthesized data. These procedures were repeated 100 times independently for each of the following scenarios:

- x_1 follows a standard normal, x_2 is a binary variable with $p = 0.5$, z_1 is a random binary variable with $p = 0.5$, $(0,1)$ and z_2 is a random uniform variable on $(0,1)$. We chose the dispersion parameter, α, to be equal to 0.6.

  $\mu_i = \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})$

  $\text{logit}(p_i) = \gamma_1 z_{1i} + \gamma_2 z_{2i}$

  we have used the following parameter values:

  $\beta_0 = 1, \beta_1 = -0.5, \beta_2 = 0.5, \gamma_1 = 1.5, \gamma_2 = -0.5$

- x_1 follows a standard normal, x_2 is a binary variable with $p = 0.8$, x_3 is Poisson distributed with mean 3, z_1 is a random uniform variable on $(0,1)$, z_2 follows a binary distribution with $p = 0.4$. We chose the dispersion parameter, α, to be equal to 0.3.

  $\mu_i = \exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i})$

  $\text{logit}(p_i) = \gamma_1 z_{1i} + \gamma_2 z_{2i}$

  we have used the following parameter values:

  $\beta_0 = -0.6,\ \beta_1 = 1.1, \beta_2 = 0.3, \gamma_1 = -0.5, \gamma_2 = 1.5$

The results are summarized in Tables (2-5) and (2-6). The generalized Poisson model provided good estimates of all coefficients except for the intercept which were estimated with a bias of size 0.492. Based on this model, we would calculate the baseline prevalence rate to be exp (0.508) =1.661, when the true value is exp (1) =2.718. On the other hand, the generalized Poisson model for underreported counts were able to accurately estimate all coefficients and provide further insight about the underlying factors that contribute to underreporting. This suggests that, when underreporting is present, the generalized Poisson regression might lead to misleading inferences due to biased estimates it produces for the model's intercept. In a similar manner, we can conclude that the estimates from undercount generalized Poisson model are more accurate compared to the ones provided by the standard generalized Poisson regression.

Table 2.5 Simulation results for generalized Poisson model

First simulation scenario				Second simulation scenario			
True value	Mean	SD	Bias	True Value	Mean	SD	Bias
$\beta_0 = 1$	0.508	0.026	0.492	$\beta_1 = 0.7$	0.649	0.007	0.051
$\beta_1 = -0.5$	-0.490	0.012	-0.01	$\beta_2 = -0.2$	-0.253	0.018	0.053
$\beta_2 = 0.5$	0.494	0.025	0.006	$\beta_3 = 0.6$	0.535	0.002	0.065
$\alpha = 0.6$	0.613	0.005	-0.013	$\alpha = 0.4$	0.514	0.005	-0.114

Table 2.6 Simulation results for generalized Poisson undercount model

First simulation scenario				Second simulation scenario			
True value	Mean	SD	Bias	True Value	Mean	SD	Bias
$\beta_0 = 1$	0.994	0.037	0.006	$\beta_1 = 0.7$	0.699	0.006	0.001
$\beta_1 = -0.5$	-0.501	0.012	0.001	$\beta_2 = -0.2$	-0.200	0.013	0.000
$\beta_2 = 0.5$	0.504	0.025	-0.004	$\beta_3 = 0.6$	0.600	0.002	0.000
$\gamma_1 = 1.5$	1.512	0.129	-0.012	$\gamma_1 = -0.4$	-0.401	0.027	0.001
$\gamma_2 = -0.5$	-0.500	0.081	0.000	$\gamma_2 = 2$	2.004	0.084	-0.004
$\alpha = 0.6$	0.598	0.006	0.002	$\alpha = 0.4$	0.399	0.005	0.001

CHAPTER 3
REGRESSION MODELS FOR MIS-REPORTED COUNTS

Underreporting is a widespread problem especially in survey research when respondents might provide inaccurate information either purposefully or due to forgetting and memory failure (Sellers, 2011). Since basing the analysis on inaccurate information could have detrimental effects on associated inferences, several methodological approaches have been proposed to adjust for underreporting in count data.

While such a framework is useful in capturing true number of events when only a fraction is reported, it is important to develop a more flexible model that covers a broader range of bias associated with misreporting (either under- or overreporting). To address this, Li et al. (Li, Trivedi, & Guo, 2003) considered a structural approach to model a potentially misreported count. Specifically, they assumed for the true count variable to follow a negative binomial regression while the reported count variable follows a Poisson regression. They estimated the model parameters through simulated maximum likelihood method. Pararai et al. (Pararai et al., 2010) used a similar approach but considered a generalized Poisson regression for the true counts instead of a negative binomial. They chose the standard maximum likelihood methods for their estimation process.

We extended the ideas suggested by Li et al. (Li et al., 2003) and Pararai et al. (Pararai et al., 2010) and developed two mixture models to explain misreported counts

when the true but unobserved counts follow either a Poisson or a generalized Poisson distribution. While the latter might seem similar to the generalized Poisson mixture model proposed by Pararai et al. (Pararai et al., 2010), we used simulated maximum likelihood method instead of the standard ML procedure for estimating model parameters.

Upcoming next, we will first discuss the simulated maximum likelihood method in Section 3.1 and then will introduce Poisson model for misreported counts and generalized Poisson model for misreported counts in sections 3.2 and 3.3 respectively.

3.1 SIMULATED MAXIMUM LIKELIHOOD ESTIMATION

Simulation based methods have played an increasingly large role in various fields such as statistics and econometrics. Despite the fact that they are computationally expensive, the recent improvements in computer hardware and software have made simulation methods even more popular (Greene, 2003). The payoff has been in the form of methods for modeling complicated processes and solving estimation problems that did not have an analytic solution. Simulation methods are mainly used for explaining characteristics of random variables including test statistics, estimators or functions of estimators. When the statistical properties of such variables cannot be derived explicitly, it is often possible to infer them through sampling from their distribution (Smita, 2009). In more recent years simulation methods have been applied not only to make inferences about an estimator but also to ease the estimation process itself.

Sometimes the likelihood function of the model involves complicated integrals that do not have a closed form solution. Generally, it is a result of missing an endogenous

variable or partially observing one that a non-tractable integral appears (Arias & Cox, 1999). In such cases simulation methods can be used to evaluate the unsolvable model within acceptable degrees of approximation. The idea behind this is that the integrals of interest are probabilities of a specific event in a random process. So, by simulating that random process, the empirical probability of the event can be used as an approximation to the value of the intractable integral we are interested in (Lerman & Manski, 1981). This idea has been labeled as probability simulation method in the literature. In fact, the method of simulated likelihood is essentially a classical sampling theory rather than being a tool for computing high dimensional integrals.

Gouriéroux and Manfort have provided detailed discussion of the SML method in their book (Gourieroux, Gourieroux, Monfort, & Monfort, 1996). Here, I briefly review the method so that I can later use it for estimation purposes.

To illustrate and begin the development of simulated maximum likelihood (SML) estimator, we consider θ to be the parameter of interest which we wish to estimate through standard ML

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \sum_{i=1}^{n} \log f(y_i|x_i, \theta) \tag{3.1}$$

Suppose $f(y_i|x_i, \theta)$, the conditional pdf of Y, has an intractable form. Suppose we have at our disposal an unbiased simulator $\tilde{f}(y_i, x_i, u, \theta)$ such that

$$E_u\left(\tilde{f}(y_i,\ x_i,\ u; \theta)|x_i, y_i\right) = f(y_i|x_i; \theta) \tag{3.2}$$

where u is an auxiliary variable with a known distribution. Then for each $i, i = 1, \dots, n$, one may have S independent random draws $u_i^s, s = 1, \dots, S$ from a known density by

which u_i^s are distributed. An SML estimator of θ can be defined as

$$\hat{\theta}_{s,n} = arg\, max_\theta \sum_{i=1}^{n} log\left[\frac{1}{S}\sum_{s=1}^{S} \tilde{f}(y_i,\, x_i,\, u_i^s;\, \theta)\right] \tag{3.3}$$

The asymptotic properties of the SML estimator can be evaluated under two circumstances

i. When S, the number of random draws from the auxiliary variable u, is fixed. Let $S = 1$, then if n goes to infinity, we have

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \log \tilde{f}(y_i,\, x_i,\, u_i^s;\, \theta) =$$

$$E \int \log \tilde{f}(y,\, x,\, u;\, \theta)\, g(u) du \tag{3.4}$$

where g is the pdf of u. Based on the definition, $\tilde{f}(y,\, x,\, u;\, \theta)$ is an unbiased simulator of f. However, in general, $\log \tilde{f}(y,\, x,\, u;\, \theta)$ is not an unbiased simulator for $\log f$. So, the result of maximizing (3.4) would not be equal to the true value of the parameter θ which is the solution to

$$\max_\theta E \log f(y|x;\theta)$$

Thus, when S is fixed, the SML estimator is not consistent.

ii. When S and n both goes to infinity

$$\lim_{S,n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \log\left[\frac{1}{S}\sum_{s=1}^{S} \tilde{f}(y_i,\, x_i,\, u_i^s;\, \theta)\right]$$

$$= \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \log\left[\int \tilde{f}(y_i,\, x_i,\, u;\, \theta) g(u) du\right]$$

$$= E \log\left[\int \tilde{f}(y_i,\, x_i,\, u;\, \theta) g(u) du\right]$$

$$= E \log f(y_i,\ x_i;\ \theta)$$

The last equation resulted owing to the fact that $\tilde{f}$ is an unbiased simulator of f. Thus, if n and S both go to infinity the SML estimator would be consistent.

It has also been proved by Gouriéroux and Manfort (Gourieroux et al., 1996) that if $S, n \to \infty$ and $\sqrt{n}/S \to 0$, then the SML estimator is asymptotically equivalent to the ML estimator.

An important step toward getting an SML estimator is finding an unbiased simulator, $\tilde{f}$. The accomplishment of this step largely depends on the form of the function f. In situations where the conditional pdf has an integral form

$$f(y_i|x_i;\theta) = \int f^*(y_i|x_i;u;\theta)g(u)du \tag{3.5}$$

It is possible to introduce the simulator

$$\tilde{f}(y,x,u;\theta) = f^*(y|x;u;\theta)$$

Where u has a distribution with pdf g.

In cases where drawing from the target distribution $g(u)$ appears to be impossible or with hardship, importance sampling can be used to draw from a more convenient distribution. Let φ be an importance function with the same support as g, such that

$$\varphi > 0,\ \int \varphi(u)du = 1$$

Without loss of generality we can rewrite f as

$$f(y_i|x_i;\theta) = \int f^*(y_i|x_i;u;\theta)g(u)\frac{\varphi(u)}{\varphi(u)}du =$$

$$E_u\left[f^*(y|x;u;\theta)\frac{g(u)}{\varphi(u)}\right] \tag{3.6}$$

Now we can introduce the simulator

$$\tilde{f}(y, x, u; \theta) = f^*(y|x; u; \theta)\frac{g(u)}{\varphi(u)} \tag{3.7}$$

where u has a distribution with pdf φ.

3.2 POISSON MODEL FOR MISREPORTED COUNTS

One of the concerns in regression modeling including count outcomes is getting biased estimates for the parameters. That concern would be even greater when the count being studied are likely to be mismeasured or misreported (Bennett, 2011). Misreporting can emerge in the form of counts being inflated (overreporting) or lessened (underreporting). Ignoring the misreporting pattern in count data can give rise to bias in the estimation of model parameters. While considerable effort has been made to promote count models in a way they can capture underreporting, less attention has been paid to developing a more flexible class of models that can adjust for a broader range of bias in reported counts.

In this section, we introduce a Poisson model that can be used in the presence of either underreporting, overreporting or even correctly reporting.

The main assumption is that the number of counts we observe is the result of two consecutive processes. The first process is the one taking care of the accurate counts while the second one is responsible for introducing underreporting or overreporting.

The construction of the model can be described as follows:

i. The true number of events, $y_i^*, i = 1, \dots, n$, follows a Poisson distribution with mean λ_i

$$f(y_i^*) = \frac{e^{-\lambda_i}\,\lambda_i^{y_i^*}}{y_i^*!}, y_i^* = 0,1,2,\ldots \tag{3.8}$$

$$\lambda_i = \exp(\boldsymbol{x}_i\boldsymbol{\gamma}) \tag{3.9}$$

where $\boldsymbol{x}$ is a row-vector of explanatory variables containing information about individual's characteristics and $\boldsymbol{\gamma}$ is the vector of unknown parameters related to marginal means of the true but unobserved counts.

ii. If y_i^*, the true number of events, is zero, the observed counts are either correctly reported as zero or they are overreported to some positive numbers. Since Poisson distribution is a common pattern for non-negative valued data, we assumed for the observed counts, y_i, to be Poisson distributed with mean μ_i, given that the actual number of events is zero.

$$f(y_i|y_i^* = 0) = \frac{e^{-\mu_i}\,\mu_i^{y_i}}{y_i!}, y_i = 0,1,2,\ldots \tag{3.10}$$

$$\mu_i = \exp(\boldsymbol{z}_i\boldsymbol{\delta}) \tag{3.11}$$

where $\boldsymbol{z}$ is a set of covariates believed to be in relation with the conditional mean of the observed counts and $\boldsymbol{\delta}$ is a vector of unknown parameters. While the value of μ can be any non-negative integer, we expect it to be zero on average.

iii. If the actual number of events is y_i^* where $y_i^* > 0$, then the reported counts can be either greater than y_i^* (overreporting) or lower than y_i^* (underreporting). To model such a bias we assume for the observed counts y_i, to follow Poisson distribution with mean $y_i^*\eta_i$

$$f(y_i|y_i^* > 0) = \frac{e^{-y_i^*\eta_i}\,(y_i^*\eta_i)^{y_i}}{y_i!}\,, y_i = 0,1,2,\ldots \tag{3.12}$$

$$\eta_i = \exp(\boldsymbol{z}_i\boldsymbol{\beta}) \tag{3.13}$$

Where $\boldsymbol{z}$ is some exploratory variables related to the conditional mean of the observed counts, given the true counts and $\boldsymbol{\beta}$ is a vector of unknown parameters. If $\eta = 1$, the counts are correctly reported. When $\eta > 1$, the events reported are higher than the actual counts and when $\eta < 1$, only a fraction of the actual events are reported.

Thus, the structure of the model allows us accommodate both underreporting and overreporting. The aim is to use the information from the observed number of events y and external variables $\boldsymbol{x}$ and $\boldsymbol{z}$ to estimate the parameters $\boldsymbol{\gamma}, \boldsymbol{\delta}$ and $\boldsymbol{\beta}$. Combining the assumptions (i)-(iii) we can write the probability mass function of the observed counts as

$$f(y_i|\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\beta}) = \sum_{y_i^*=0}^{\infty} Pr(y_i|y_i^*,\ \boldsymbol{z}_i, \boldsymbol{\delta}, \boldsymbol{\beta})\, Pr(y_i^*|\boldsymbol{x}_i,\ \boldsymbol{\gamma})$$

$$= Pr(y_i|y_i^* = 0,\ \boldsymbol{z}_i, \boldsymbol{\delta}, \boldsymbol{\beta})\, Pr(y_i^* = 0|\boldsymbol{x}_i, \boldsymbol{\gamma}) + \sum_{y_i^*=1}^{\infty} Pr(y_i|y_i^*,\ \boldsymbol{z}_i, \boldsymbol{\delta}, \boldsymbol{\beta})\, Pr(y_i^*|\boldsymbol{x}_i,\ \boldsymbol{\gamma})$$

$$= \frac{e^{-\mu_i}\,{\mu_i}^{y_i}}{y_i\,!} \cdot e^{-\lambda_i} + \sum_{y_i^*=1}^{\infty} \frac{e^{-y_i^*\eta_i}\,(y_i^*\eta_i)^{y_i}}{y_i\,!} \cdot \frac{e^{-\lambda_i}\,{\lambda_i}^{y_i^*}}{y_i^*\,!} \tag{3.14}$$

Now that we have the likelihood function for the i[th] individual at hand, we can use some maximization method to estimate the parameters of interest. However, due to the presence of the infinite series in (3.14), the likelihood function does not have a closed form solution. As a possible remedy, one might consider replacing the upper limit of the sum with a relatively large cut point assuming that the remainder of the series becomes

negligible from that point forward. According to Li et al. (2003), "such a method, however, results in inconsistent estimates due to the truncation of the true likelihood function. In addition, it is an ad hoc method of choosing the truncation point".

As an alternative, we use simulated maximum likelihood discussed in Section 3.1 for estimating the parameters of our model. To that goal, the first step would be finding an unbiased simulator for the likelihood function (3.14). Introducing an importance function φ to the likelihood, we can rewrite (3.14) as

$$f(y_i|\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\beta}) = \frac{e^{-\mu_i}\,{\mu_i}^{y_i}}{y_i\,!} \cdot e^{-\lambda_i} + \sum_{y_i^*=1}^{\infty} \frac{e^{-y_i^*\eta_i}\,(y_i^*\eta_i)^{y_i}}{y_i\,!} \cdot \frac{e^{-\lambda_i}\,{\lambda_i}^{y_i^*}}{y_i^*\,!}$$

$$= \frac{e^{-\mu_i}\,{\mu_i}^{y_i}}{y_i\,!} \cdot e^{-\lambda_i} + \sum_{y_i^*=1}^{\infty} \frac{Pr(y_i|y_i^*,\, \boldsymbol{z}_i, \boldsymbol{\beta})\, Pr(y_i^*|\boldsymbol{x}_i, \boldsymbol{\gamma})\, \varphi(y_i^*|\boldsymbol{x}_i)}{\varphi(y_i^*|\boldsymbol{x}_i)}$$

$$= E_{y_i^*}\left[\frac{e^{-\mu_i}\,{\mu_i}^{y_i}}{y_i\,!} \cdot e^{-\lambda_i} + \frac{Pr(y_i|y_i^*,\, \boldsymbol{z}_i, \boldsymbol{\beta})\, Pr(y_i^*|\boldsymbol{x}_i, \boldsymbol{\gamma})}{\varphi(y_i^*|\boldsymbol{x}_i)} \right] \tag{3.15}$$

Therefore, an unbiased simulator for f can be chosen as

$$\tilde{f}(y_i|\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\beta}) = \frac{e^{-\mu_i}\,{\mu_i}^{y_i}}{y_i\,!} \cdot e^{-\lambda_i} + \frac{Pr(y_i|u,\, \boldsymbol{z}_i, \boldsymbol{\beta})\, Pr(u|\boldsymbol{x}_i, \boldsymbol{\gamma})}{\varphi(u|\boldsymbol{x}_i)} \tag{3.16}$$

where u is an auxiliary variable that only takes integer values greater than or equal to one and its probability mass function is represented as $\varphi(u|\boldsymbol{x}_i)$. Any distribution that satisfies

$$\varphi(u_i|\boldsymbol{x}_i) = \frac{e^{-\Delta_i}{\Delta_i}^{u_i}}{u_i!\,(1 - e^{-\Delta_i})}\,, u_i = 1,2,3, \ldots \tag{3.17}$$

this condition can serve as the importance function. We selected u from a zero truncated Poisson distribution with mean Δ

where Δ can be estimated by fitting a Poisson model to the non-zero observations in y, using the explanatory variables $\boldsymbol{x}$.

Thus, we can rewrite the likelihood function (3.14) using the suggested $\tilde{f}$ as

$$L(y_i) \approx \frac{e^{-\mu_i}\,{\mu_i}^{y_i}}{y_i\,!} \cdot e^{-\lambda_i} + \frac{1}{S}\sum_{s=1}^{S} \frac{\frac{e^{-\eta_i u_i^s}\,(\eta_i u_i^s)^{y_i}}{y_i\,!} \cdot \frac{e^{-\lambda_i}\,{\lambda_i}^{u_i^s}}{u_i^s\,!}}{\frac{e^{-\Delta_i}\,{\Delta_i}^{u_i^s}}{u_i^s\,!\,(1-e^{-\Delta_i})}} \tag{3.18}$$

In summary, simulating the likelihood function (3.14) consists of the following steps

i. Regressing the non-zero observations in y on $\boldsymbol{x}$ through Poisson model and estimating Δ

ii. For each y_i, getting S random draws, u_i^s, from a zero truncated Poisson distribution with mean $\widehat{\Delta}$

iii. For each u_i^s, evaluating the summand in (3.19) and averaging over those values

iv. Calculating the likelihood

Having simulated the likelihood function, we can estimate the parameters $\boldsymbol{\gamma}, \boldsymbol{\delta}$ and $\boldsymbol{\beta}$ through maximization methods.

3.3 SIMULATION STUDY FOR POISSON MODEL FOR MISREPORTED COUNTS

Unlike the simulation studies discussed in previous sections, here I simulated just one dataset. The reason was that getting SML estimates requires generating random draws from the target distribution for each observation. To avoid the long processing time, I chose a simulation size of 1.

In order to assess the performance of the model we synthesized a data set of size

10,000 observations with variables (x_1, x_2) defining the true counts and (z_1, z_2) relating to misreporting. x_1 were generated from a uniform distribution and x_2 were from binomial (4,0.2). z_1 was assumed to be from a Bernoulli distribution with p=0.3 and z_2 were generated from standard normal distribution. The variable containing the true counts, y^*, were produced by generating random numbers from a Poisson distribution with mean $\exp(\gamma_0 + \gamma_1 x_1 + \gamma_2 x_2)$ where $\gamma_0 = 1$, $\gamma_1 = 1.5$ and $\gamma_2 = -0.5$. The observed counts, y, were then created based on the variable y^*. For those observations where the true number of events were zero, y were generated from a Poisson distribution with mean $\exp(\delta_1 z_1 + \delta_2 z_2)$ where $\delta_1 = -1.8$ and $\delta_2 = -1.1$. The observed counts, y, for the rest of the dataset were generated from Poisson distribution with mean $[y^* \times \exp(\beta_1 z_1 + \beta_2 z_2)]$ where $\beta_1 = 0.5$ and $\beta_2 = 0.2$. Once all variables were created, we fit first a standard Poisson regression model and then a Poisson regression model for underreported counts to the synthesized data. For the first simulation scenario we used full models and for the second we considered models with no intercept. The results are summarized in Tables (3-1) and (3-2).

While the main parameters were well estimated from both models, the results from the miscounted Poisson model were more accurate. We were also able to get some information about the sources of underreporting and overreporting through the miscounted model which is something that clearly the standard Poisson regression cannot provide.

The results from second simulation scenario suggests that even if we apply the misreporting model in situations where the counts are fully observed, we would still get reliable estimates. Table (3-2) shows that the parameters related to misreporting,

$\delta_1, \delta_2, \beta_1, \beta_2$, are all insignificant, due to the large estimated standard errors. Thus, although we expect to get better estimates using the standard Poisson regression when there are no reporting errors, applying the miscounted Poisson model would still provide acceptable results.

Table 3.1 Comparing standard Poisson to Poisson model for misreported counts, first simulation scenario

	Poisson model			Poisson model for misreported counts		
True value	Estimated Coefficient	Standard Error	Bias	Estimated Coefficient	Standard Error	Bias
$\gamma_0 = 1$	1.304	0.034	-0.304	1.069	0.056	-0.069
$\gamma_1 = 1.5$	1.433	0.048	0.067	1.478	0.075	0.022
$\gamma_2 = -0.5$	-0.540	0.019	0.04	-0.539	0.030	0.039
$\delta_1 = -1.8$	-	-	-	-1.363	1.031	-0.437
$\delta_2 = -1.1$	-	-	-	-0.759	0.251	-0.341
$\beta_1 = 0.5$	-	-	-	0.504	0.039	0.004
$\beta_2 = 0.2$	-	-	-	0.182	0.0201	0.018

Table 3.2 Comparing Standard Poisson to Poisson model for misreported counts, second simulation scenario

	Poisson model			Poisson model for misreported counts		
True value	Estimated Coefficient	Standard Error	Bias	Estimated Coefficient	Standard Error	Bias
$\gamma_1 = 1.5$	1.629	0.012	-0.129	1.511	0.021	-0.011
$\gamma_2 = -0.5$	-0.199	0.008	-0.301	-0.501	0.017	0.001
$\delta_1 = -1.8$	-	-		-1.743	0.111	-0.057
$\delta_2 = -1.1$	-	-		-1.110	0.012	0.01
$\beta_1 = 0.5$	-	-		0.464	0.019	0.036
$\beta_2 = 0.2$	-	-		0.197	0.010	0.003

3.4 GENERALIZED POISSON MODEL FOR MISREPORTED COUNTS

Negative binomial and generalized Poisson regressions are popular alternatives to Poisson regression. In Section 3.1 we introduced the Poisson model for misreported counts. Li et al. (2003) has also suggested a negative Binomial model that can accommodate both under/overreported events. Now it is natural to derive an extension to to the generalized Poisson model so that it can adjust for misreported counts along the same way as Poisson and negative binomial regression.

Similar to Poisson model for underreported counts, the construction of the generalized Poisson model for underreporting is based on the following assumptions

iv. The true number of events, $y_i^*, i = 1, \dots, n$, follows a generalized Poisson distribution with mean λ_i and dispersion parameter α

$$f(y_i^*) = \left(\frac{\lambda_i}{1+\alpha\lambda_i}\right)^{y_i^*} \frac{(1+\alpha y_i^*)^{y_i^*-1}}{y_i^*!} \exp\left[\frac{-\lambda_i(1+\alpha y_i^*)}{1+\alpha\lambda_i}\right], y_i^* = 0,1,\dots \quad (3.20)$$

$$\lambda_i = \exp(\boldsymbol{x}_i\boldsymbol{\gamma}) \quad (3.21)$$

where $\boldsymbol{x}$ is some explanatory variables containing information about individual's characteristics and $\boldsymbol{\gamma}$ is the vector of unknown parameters related to marginal means of the true but unobserved counts.

v. If y_i^*, the true number of events, is zero, the observed counts are either correctly reported as zero or they are overreported to some positive numbers. Since Poisson distribution is a common pattern for non-negative valued data, we assumed for the observed counts, y_i, to be Poisson distributed with mean μ_i, given that the actual number of events is zero.

$$f(y_i|y_i^* = 0) = \frac{e^{-\mu_i}\,\mu_i^{y_i}}{y_i!}, y_i = 0,1,2, \ldots \quad (3.22)$$

$$\mu_i = \exp(\boldsymbol{z}_i\boldsymbol{\delta}) \quad (3.23)$$

where $\boldsymbol{z}$ is a set of covariates believed to be in relation with conditional mean of the observed counts and $\boldsymbol{\delta}$ is a vector of unknown parameters. While the value of μ can be any non-negative integer, we expect it to be zero on average.

vi. If the actual number of events is y_i^* where $y_i^* > 0$, then the reported counts can be either greater than y_i^* (overreporting) or lower than y_i^* (underreporting). To model such a bias we assume for the observed counts y_i, to follow Poisson distribution with mean $y_i^*\eta_i$

$$f(y_i|y_i^* > 0) = \frac{e^{-y_i^*\eta_i}\,(y_i^*\eta_i)^{y_i}}{y_i!}, y_i = 0,1,2, \ldots \quad (3.24)$$

$$\eta_i = \exp(\boldsymbol{z}_i\boldsymbol{\beta}) \quad (3.25)$$

Where $\boldsymbol{z}$ is some exploratory variables related to the conditional mean of the observed counts, given the true counts and $\boldsymbol{\beta}$ is a vector of unknown parameters. If $\eta = 1$, the counts are correctly reported. When $\eta > 1$, the events reported are higher than the actual counts and when $\eta < 1$, only a fraction of the actual events are reported.

Thus, the structure of the model let us to accommodate for both underreporting and overreporting. The goal is to estimate the parameters $\boldsymbol{\gamma}, \boldsymbol{\delta}$ and $\boldsymbol{\beta}$, using the information from the observed number of events y and external variables $\boldsymbol{x}$ and $\boldsymbol{z}$. Combining the assumptions (i)-(iii) we can write the probability mass function of the observed counts as

$$f(y_i|\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\beta}) = \sum_{y_i^*=0}^{\infty} Pr(y_i|y_i^*,\ \boldsymbol{z}_i, \boldsymbol{\delta}, \boldsymbol{\beta})\, Pr(y_i^*|\boldsymbol{x}_i,\ \boldsymbol{\gamma})$$

$$= Pr(y_i|y_i^* = 0,\ \boldsymbol{z}_i, \boldsymbol{\delta}, \boldsymbol{\beta})\, Pr(y_i^* = 0|\boldsymbol{x}_i, \boldsymbol{\gamma}) + \sum_{y_i^*=1}^{\infty} Pr(y_i|y_i^*,\ \boldsymbol{z}_i, \boldsymbol{\delta}, \boldsymbol{\beta})\, Pr(y_i^*|\boldsymbol{x}_i,\ \boldsymbol{\gamma})$$

$$= \frac{e^{-\mu_i}\, {\mu_i}^{y_i}}{y_i\,!} \left(\frac{-\lambda_i}{1+\alpha\lambda_i}\right)$$

$$+ \sum_{y_i^*=1}^{\infty} \frac{e^{-y_i^*\eta_i}\, (y_i^*\eta_i)^{y_i}}{y_i\,!} \left(\frac{\lambda_i}{1+\alpha\lambda_i}\right)^{y_i^*} \frac{(1+\alpha y_i^*)^{y_i^*-1}}{y_i^*!} \exp\left[\frac{-\lambda_i(1+\alpha y_i^*)}{1+\alpha\lambda_i}\right] \tag{3.26}$$

Generally, once we develop the likelihood function, we would use some maximization method to estimate the parameters of interest. However, similar to section 3.2, due to the presence of infinite series in (3.14), the likelihood function does not have a closed form solution.

As an alternative, we use simulated maximum likelihood discussed in section 3.1 for estimating the parameters $\boldsymbol{\gamma}, \boldsymbol{\delta}$ and $\boldsymbol{\beta}$. To reach that goal, the first step is to find an unbiased simulator for the likelihood function (3.26). If we introduce an importance function φ to the likelihood, we can rewrite (3.26) as

$$f(y_i|\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\beta}) = \frac{e^{-\mu_i}\, {\mu_i}^{y_i}}{y_i\,!} \left(\frac{-\lambda_i}{1+\alpha\lambda_i}\right) +$$

$$\sum_{y_i^*=1}^{\infty} \frac{e^{-y_i^*\eta_i}\, (y_i^*\eta_i)^{y_i}}{y_i\,!} \left(\frac{\lambda_i}{1+\alpha\lambda_i}\right)^{y_i^*} \frac{(1+\alpha y_i^*)^{y_i^*-1}}{y_i^*!} \exp\left[\frac{-\lambda_i(1+\alpha y_i^*)}{1+\alpha\lambda_i}\right]$$

$$= \frac{e^{-\mu_i}\, {\mu_i}^{y_i}}{y_i\,!} \left(\frac{-\lambda_i}{1+\alpha\lambda_i}\right) + \sum_{y_i^*=1}^{\infty} \frac{Pr(y_i|y_i^*,\ \boldsymbol{z}_i, \boldsymbol{\beta})\, Pr(y_i^*|\boldsymbol{x}_i, \boldsymbol{\gamma})\, \varphi(y_i^*|\boldsymbol{x}_i)}{\varphi(y_i^*|\boldsymbol{x}_i)}$$

$$= E_{y_i^*}\left[\frac{e^{-\mu_i}\,{\mu_i}^{y_i}}{y_i\,!}\left(\frac{-\lambda_i}{1+\alpha\lambda_i}\right)+\frac{Pr(y_i|y_i^*,\,\boldsymbol{z}_i,\boldsymbol{\beta})\,Pr(y_i^*|\boldsymbol{x}_i,\boldsymbol{\gamma})}{\varphi(y_i^*|\boldsymbol{x}_i)}\right] \tag{3.27}$$

Therefore, an unbiased simulator for f can be chosen as

$$\tilde{f}(y_i|\boldsymbol{x}_i,\boldsymbol{z}_i,\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\beta}) = \frac{e^{-\mu_i}\,{\mu_i}^{y_i}}{y_i\,!}\left(\frac{-\lambda_i}{1+\alpha\lambda_i}\right)+\frac{\Pr(y_i|u,\,\boldsymbol{z}_i,\boldsymbol{\beta})\,\Pr(u|\boldsymbol{x}_i,\boldsymbol{\gamma})}{\varphi(u|\boldsymbol{x}_i)} \tag{3.28}$$

where u is an auxiliary variable that only takes integer values greater than or equal to one and its probability mass function is represented as $\varphi(u|\boldsymbol{x}_i)$. Any distribution that satisfy this condition can serve as the importance function. We selected u from a zero truncated generalized Poisson distribution with mean Δ and dispersion parameter ε

$$\varphi(u_i|\boldsymbol{x}_i) = \left(\frac{\Delta_i}{1+\varepsilon\Delta_i}\right)^{u_i}\times \frac{(1+\varepsilon u_i)^{u_i-1}}{u_i!\,\left[1-exp\left(\frac{-\Delta_i}{1+\varepsilon\Delta_i}\right)\right]}\exp\left[\frac{-\Delta_i(1+\varepsilon u_i)}{1+\varepsilon\Delta_i}\right],\ u_i = 1,2,3,\ldots \tag{3.29}$$

where Δ can be estimated by fitting a generalized Poisson model to the non-zero observations in y, using the explanatory variables $\boldsymbol{x}$. Thus, we can rewrite the likelihood function (3.26) using the suggested $\tilde{f}$ as

$$L(y_i) \approx \frac{e^{-\mu_i}\,{\mu_i}^{y_i}}{y_i\,!}\left(\frac{-\lambda_i}{1+\alpha\lambda_i}\right)+ \frac{1}{S}\sum_{s=1}^{S}\frac{\frac{e^{-\eta_i u_i^s}\,(\eta_i u_i^s)^{y_i}}{y_i\,!}\cdot\left(\frac{\lambda_i}{1+\alpha\lambda_i}\right)^{u_i^s}\frac{(1+\alpha u_i^s)^{u_i^s-1}}{u_i^s!}\exp\left[\frac{-\lambda_i(1+\alpha u_i^s)}{1+\alpha\lambda_i}\right]}{\left(\frac{\Delta_i}{1+\varepsilon\Delta_i}\right)^{u_i}\cdot\frac{(1+\varepsilon u_i)^{u_i-1}}{u_i!\,\left[1-exp\left(\frac{-\Delta_i}{1+\varepsilon\Delta_i}\right)\right]}\exp\left[\frac{-\Delta_i(1+\varepsilon u_i)}{1+\varepsilon\Delta_i}\right]} \tag{3.30}$$

In summary, simulating the likelihood function (3.14) consists of the following steps

v. Regressing the non-zero observations in y on $\boldsymbol{x}$ through generalized Poisson model and estimating Δ and ε

vi. For each y_i, getting S random draws, u_i^s, from a zero truncated generalized Poisson distribution with mean $\widehat{\Delta}$ and dispersion parameter $\hat{\varepsilon}$

vii. For each u_i^s, evaluating the summand in (3.31) and averaging over those values

viii. Calculating the likelihood

Having simulated the likelihood function, we can estimate the parameters $\boldsymbol{\gamma}, \boldsymbol{\delta}$ and $\boldsymbol{\beta}$ through maximization methods.

3.4 SIMULATION STUDY FOR GENERALIZED POISSON MODEL FOR MISREPORTED COUNTS

In order to assess the performance of the model I synthesized a data set of size 1000 observations with variables (x_1, x_2) defining the true counts and (z_1, z_2) relating to misreporting. x_1 were generated from a uniform distribution on the interval (0,3) and x_2 were from binomial (4,0.2). z_1 is assumed to be from a uniform distribution on the interval (0,1) and z_2 were generated from a binomial distribution with n=5 and p=0.2. The variable containing the true counts, y^* were produced by generating random numbers from a generalized Poisson distribution with mean $\exp(\gamma_0 + \gamma_1 x_1 + \gamma_2 x_2)$ and dispersion parameter α where $\gamma_0 = 0.2$, $\gamma_1 = 0.4$, $\gamma_2 = -1.5$ and $\alpha = 0.45$. The observed counts, y, were then created based on the variable y^*. For those observations where the true number of events were zero, y were generated from a Poisson distribution with mean $\exp(\delta_1 z_1 + \delta_2 z_2)$ where $\delta_1 = -2.7$ and $\delta_2 = 0.4$. The observed counts, y, for

the rest of the dataset were generated from Poisson distribution with mean $[y^* \times \exp(\beta_1 z_1 + \beta_2 z_2)]$ where $\beta_1 = 0.1$ and $\beta_2 = 0.3$. Once all variables were created, we fit first a standard generalized Poisson regression (GPR) model and then a generalized Poisson regression model for misreported counts (GPRM) to the synthesized data.

For getting the SML estimates, we drew 100 observations from zero truncated generalized Poisson distribution with mean and dispersion parameter estimated from the naive model.

For the second simulation scenario, we considered a situation where the counts were fully observed. Our aim was to observe the performance of the model when the number of events were all correctly reported, i.e. $y = y^*$. We used the same variable layouts for this part. The results are summarized in Tables (3-3) and (3-4).

Comparing the true value of the parameters with those estimated from the GP model reveals that when there exist patterns of overreporting or underreporting in the data, applying the GP model might result in biased estimates.

Unlike the GP mode, the estimates from the generalized Poisson model for misreported counts were more accurate. The GP model for misreporting were also able to identify the variables contributing to underreporting and overreporting separately.

For the second simulation scenario, where the counts were fully observed, both models provided good estimates of the main parameters, γ_0, γ_1 and γ_2. Moreover, the parameters explaining the magnitude of misreporting introduced by variables, (z_1, z_2), are either very small or insignificant due to the large standard errors. This suggests that, if

Table 3.3 Comparing generalized Poisson to generalized Poisson model for underreported counts, first simulation scenario

	Generalized Poisson model			Generalized Poisson model for misreported counts		
True value	Estimated Coefficient	Standard Error	Bias	Estimated Coefficient	Standard Error	Bias
$\gamma_0 = 0.2$	0.744	0.099	-0.544	0.127	0.163	0.073
$\gamma_1 = 0.4$	0.257	0.046	0.143	0.457	0.066	-0.057
$\gamma_2 = -1.5$	-0.636	0.058	-0.864	-1.388	0.138	-0.112
$\alpha = 0.45$	0.581	0.019	-0.131	0.465	0.038	-0.015
$\delta_1 = -2.7$	-	-	-	-2.975	0.314	0.275
$\delta_2 = 0.4$	-	-	-	0.388	0.061	0.012
$\beta_1 = 0.1$	-	-	-	-0.170	0.150	0.27
$\beta_2 = 0.3$	-	-	-	0.387	0.055	-0.087

Table 3.4 Comparing generalized Poisson to generalized Poisson model for underreported counts, second simulation scenario

	Generalized Poisson model			Generalized Poisson model for misreported counts		
True value	Estimated Coefficient	Standard Error	Bias	Estimated Coefficient	Standard Error	Bias
$\gamma_0 = 0.2$	0.119	0.120	0.081	0.143	0.057	0.057
$\gamma_1 = 0.4$	0.430	0.056	-0.030	0.430	0.136	-0.030
$\gamma_2 = -1.5$	-1.519	0.098	0.019	-1.504	0.100	0.004
$\alpha = 0.45$	0.474	0.025	-0.024	0.310	0.044	0.140
$\delta_1 = -2.7$	-	-	-	-24.405	13.124	21.705
$\delta_2 = 0.4$	-	-	-	-19.79	8995.6	20.190
$\beta_1 = 0.1$	-	-	-	-0.192	0.139	0.292
$\beta_2 = 0.3$	-	-	-	0.068	0.050	0.232

we do not know whether there are reporting biases in the data and still use the GP misreporting model, the results would still not be off. Clearly, however, we expect for the GP model to provide better estimates in that situation.

CHAPTER 4
APPLICATION TO EBAN STUDY

4.1 INTRODUCTION

Around 1.1 million people are living with HIV in the United States (CDC, 2018b). While the size of the epidemic is relatively small compared to the country's population, the disproportionate burden of the disease among certain minority groups has been of concern for years (Aral, Adimora, & Fenton, 2008; Control & Prevention, 2011). According to CDC, African Americans have the highest rate of HIV when compared to other races. In 2016, African Americans accounted for 44% of HIV diagnoses, though they comprise 12% of the U.S. population.

"A number of challenges contribute to the higher rates of HIV infection among African Americans. The greater number of people living with HIV (prevalence) in African American communities and the tendency for African Americans to have sex with partners of the same race/ethnicity mean that African Americans face a greater risk of HIV infection. Some African American communities also experience higher rates of other sexually transmitted diseases (STDs) than other racial/ethnic communities in the United States. Having another STD can significantly increase a person's chance of getting or transmitting HIV"(CDC, 2018a).

In an attempt to determine whether an intervention method could be effective in reducing high risk behaviors among African Americans, an RCT with a focus on African

American HIV serodiscordant heterosexual couples were conducted in 2007. The study individuals were recruited from 4 sites: Atlanta, Georgia; Los Angeles, California; New York, New York; and Philadelphia, Pennsylvania. Couples were eligible if they were both at least 18 years old and were aware of their partner's HIV serostatus. Once eligibility were confirmed, couples were allocated to one of 2 interventions, the Eban HIV/STD risk reduction or the health promotion comparison. Data were collected at 4 time points, baseline, right after intervention, 6 months post intervention and 12 months post intervention. The detailed description of the study can be found elsewhere (El-Bassel et al., 2010; Syndromes, 2008).

In this dissertation, I am going to use responses 6 months after the intervention. The primary outcome is whether the couple had unprotected sexual activity during past 3 months. Both partners responded to this question during an interview but here I am considering the male's participant responses. We are interested to see if the intervention group has lower rate of unprotected intercourse acts and if there exists a pattern of underreporting/overreporting in individuals' responses.

4.2 METHODS

For the first round of analysis we use the standard count models (Poisson, negative binomial and generalized Poisson regression) to get an estimate of the relationship between the number of unprotected sexual activities and the covariates of interest which includes treatment group, age, marital status, living with study partner, and having multiple concurrent partners. Next, we compare the fitted models and will choose the one with the lowest values of AIC and BIC for further analysis. Then, we will develop underreporting regression and misreporting regression models, based on the

model we selected in the previous step, to investigate any potential sources of overreporting or underreporting. Finally, we will compare the results from these models and will choose one of them for interpretation purposes.

4.3 RESULTS

Of the 535 couples that were included in the study, 260 (48.59%) were allocated to the EBAN intervention group and 275 (51.44%) were allocated to health promotion intervention. The average age of male participants that were used for further analysis were 45.89 years old with standard deviation of 8.30. While only 206 (38.50%) individuals were married, the majority were living with their study partner (61.30%).

Table (4-1) shows the results of fitting Poisson, negative binomial and generalized Poisson to the EBAN data. The estimation procedure did not converge for the generalized Poisson model, so we reported the estimates achieved after 100 iterations. The AIC and BIC are the highest for Poisson regression and the lowest for negative binomial regression. Thus, in the next step we fit extensions of the negative binomial model for underreported data and misreported data to investigate the potential errors in number of individuals diagnosed with ADRD. The results are summarized in Table (4-2).

The AIC and BIC of both models are very close to each other but since they are smaller for negative binomial underreporting model, we will focus on that model for further exploration of the results.

The first panel describes the factors related to the actual number of times the participant reported having unprotected intercourse acts. The constant is estimated to be 2.78, suggesting that the baseline incidence rate is 16.11. The coefficient for being in the

EBAN intervention group is -0.684, which shows that the rate of unprotected sexual activities for those that have received behavioral interventions is 50% less than those who were part of the health promotion group.

Table 4.1 Results from fitting Poisson, negative binomial and generalized Poisson regressions

	Coeff.	Std. Err.	z	p-value
	Poisson model			
Constant	1.745	0.132	13.14	<0.001
Treatment	-0.702	0.041	-16.8	<0.001
Age	-0.017	0.002	-7.32	<0.001
Marital status	0.122	0.039	3.13	0.002
Living with study partner	1.339	0.078	17.16	<0.001
Having other concurrent partners	-0.811	0.101	-7.98	<0.001
LL	-4768.465			
AIC	9548.93			
BIC	9573.354			
	Negative binomial model			
Constant	1.467	0.823	1.78	0.075
Treatment	-0.710	0.273	-2.6	0.009
Age	-0.010	0.017	-0.58	0.565
Marital status	0.129	0.282	0.46	0.646
Living with study partner	1.319	0.335	3.93	<0.001
Having other concurrent partners	-1.070	0.426	-2.51	0.012
α	6.902			
LL	-866.009			
AIC	1746.018			
BIC	1774.514			
	Generalized Poisson model*			
Constant	6.836	-	-	-
Treatment	-0.639	0.157	-4.06	<0.001
Age	-0.015	0.009	-1.6	0.111
Marital status	0.012	0.156	0.08	0.934
Living with study partner	0.575	0.214	2.68	0.007
Having other concurrent partners	-0.285	0.281	-1.02	0.310
α	-.993			
LL	-887.389			
AIC	1786.78			
BIC	1811.204			

* Convergence was not achieved

Table 4.2 Results from fitting negative binomial model for underreporting and negative binomial model for misreporting

		Coeff.	Std. Err.	z	p-value
		Negative binomial for underreporting			
Underreporting	Constant	2.786	0.558	4.98	<0.001
	Treatment	-0.684	0.273	-2.5	0.012
	Age	-0.033	0.018	-1.85	0.064
	Marital status	0.059	0.818	0.07	0.942
	Living with study partner	2.172	0.911	2.38	0.017
	Having other concurrent partners	-1.403	0.821	-1.71	0.088
	α	6.894			
	LL	-865.851			
	AIC	1745.704			
	BIC	1774.199			
		Negative binomial model for misreporting			
$p(y\|y^* = 0)$	Constant	1.401	0.175	7.99	<0.001
	Treatment	-0.750	0.244	-3.07	0.002
	Age	-0.102	0.031	-3.28	0.001
	Marital status	0.649	1.084	0.6	0.549
	Living with study partner	1.142	1.035	1.1	0.27
	Having other concurrent partners	-18.729	10185.17	0	0.999
$p(y\|y^* = 0)$	Age	-0.008	0.005	-1.63	0.103
	Marital status	0.501	0.086	5.79	<0.001
	Living with study partner	1.077	0.225	4.77	<0.001
	Having other concurrent partners	-0.836	0.291	-2.87	0.004
	α	5.64			
	LL	-863.235			
	AIC	1748.47			
	BIC	1793.248			

The second panel explores the chances of someone reporting a smaller number when asked about the number of times he had unprotected sexual activities during last 3

months. A negative estimated coefficient for age suggests that younger people are more likely to underreport their high risk sexual behavior. Living with study partner, on the other hand seems to be highly correlated with a pattern of underreporting.

4.4 CONCLUSION

The results confirmed that behavioral intervention can reduce HIV/sexually transmitted disease (STD) risk behaviors among African American HIV serodiscordant couples. Also, our model provided a good insight into how some factors like age and living with a partner might contribute to underreporting high risk behavior.

CHAPTER 5
CONCLUSIONS AND FUTURE WORK

Data collection often involves reporting errors. Underreporting, a more common problem in counting systems, happens when the reporting of some events is not complete. As a consequence of underreporting, the mean of the observed counts is smaller than the true mean. Ignoring the underreporting pattern of count data could result in biased estimates of the effects of interest which ultimately leads to misleading inferences.

Extensions of standard count data models (Poisson, negative binomial and generalized Poisson regression) have been proposed by various authors so that the underreporting patterns can also be captured. All these models assume a mixture of binomial distribution and some other distribution for counts. Basically, the binomial model presumes that for each event a random mechanism decides whether it is reported or not.

A key assumption among the underreporting models is that the reporting probability is constant and identical for all events. However, in practice the reporting probability might change under different circumstances. Pararai et al. (2006) suggested using quasi binomial distribution II (QBD-II) instead of binomial distribution to reach that goal. They developed a generalized Poisson model applicable to underreporting events which does not rely on the constant probability assumption. Future research is

needed to derive Poisson and negative binomial models for underreported counts which also allow for a changeable reporting probability.

Although less common, overreporting is another problem that might affect counting systems. We proposed two models that are capable of capturing both underreporting and overreporting. In situations where the outcome of interest is over dispersed and also likely to be misreported, negative binomial regression for misreported counts can be used as an alternative to negative binomial regression. In other cases where the outcome of interest might be under- or overreported, and also the distribution of counts seems to be under dispersed, we can apply generalized Poisson regression instead of the standard GP model.

Both proposed models enable us to determine how individual's characteristics contribute to reporting bias. They also provide us a way to estimate the proportions of underreporting, overreporting and correctly reporting.

BIBLIOGRAPHY

Aral, S. O., Adimora, A. A., & Fenton, K. A. J. T. L. (2008). Understanding and responding to disparities in HIV and other sexually transmitted infections in African Americans. *372*(9635), 337-340.

Arias, C., & Cox, T. L. (1999). Maximum simulated likelihood: a brief introduction for practitioners.

Bae, S., Famoye, F., Wulu, J., Bartolucci, A. A., & Singh, K. P. (2005). A rich family of generalized Poisson regression models with applications. *Mathematics and Computers in Simulation, 69*(1-2), 4-11.

Bennett, M. M. (2011). *Bayesian approaches to correcting bias in epidemiological data.*

Cameron, A. C., & Johansson, P. (1997). Count data regression using series expansions: with applications. *Journal of Applied Econometrics, 12*(3), 203-223.

Cameron, A. C., & Trivedi, P. K. (2001). Essentials of count data regression. *A companion to theoretical econometrics, 331*.

Cameron, A. C., Trivedi, P. K., Milne, F., & Piggott, J. (1988). A microeconometric model of the demand for health care and health insurance in Australia. *The Review of economic studies, 55*(1), 85-106.

CDC. (2018a). HIV among African Americans. Retrieved from https://www.cdc.gov/hiv/group/racialethnic/africanamericans/index.html

CDC. (2018b). HIV in the United States: At a glance. Retrieved from https://www.cdc.gov/hiv/pdf/statistics/overview/cdc-hiv-us-ataglance.pdf

Consul, P., & Famoye, F. (1992). Generalized Poisson regression model. *Communications in Statistics-Theory and Methods, 21*(1), 89-109.

Control, C. f. D., & Prevention. (2011). Disparities in diagnoses of HIV infection between blacks/African Americans and other racial/ethnic populations--37 states, 2005-2008. *60*(4), 93.

Contzen, N., De Pasquale, S., & Mosler, H.-J. (2015). Over-reporting in handwashing self-reports: Potential explanatory factors and alternative measurements. *PloS one, 10*(8), e0136445.

Czado, C., Erhardt, V., Min, A., & Wagner, S. (2007). Zero-inflated generalized Poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcing rates. *Statistical Modelling, 7*(2), 125-153.

Del Castillo, J., & Pérez-Casany, M. (1998). Weighted Poisson distributions for overdispersion and underdispersion situations. *Annals of the Institute of Statistical Mathematics, 50*(3), 567-585.

Donaldson, S. I., & Grant-Vallone, E. J. (2002). Understanding self-report bias in organizational behavior research. *Journal of business and Psychology, 17*(2), 245-260.

El-Bassel, N., Jemmott, J. B., Landis, J. R., Pequegnat, W., Wingood, G. M., Wyatt, G. E., & Bellamy, S. L. J. A. o. i. m. (2010). National Institute of Mental Health multisite Eban HIV/STD prevention intervention for African American HIV serodiscordant couples: a cluster randomized trial. *170*(17), 1594-1601

Fader, P. S., & Hardie, B. G. (2000). A note on modelling underreported Poisson counts.

Famoye, F., & Wang, W. (2004). Censored generalized Poisson regression model. *Computational statistics & data analysis, 46*(3), 547-560.

Famoye, F., Wulu, J. T., & Singh, K. P. (2004). On the generalized Poisson regression model with an application to accident data. *Journal of Data Science, 2*(2004), 287-295.

Gourieroux, M., Gourieroux, C., Monfort, A., & Monfort, D. A. (1996). *Simulation-based econometric methods*: Oxford university press.

Greene, W. H. (2003). *Econometric analysis*: Pearson Education India.

Lerman, S., & Manski, C. (1981). On the use of simulated frequencies to approximate choice probabilities. *Structural analysis of discrete data with econometric applications, 10*, 305-319.

Li, T., Trivedi, P. K., & Guo, J. (2003). Modeling response bias in count: a structural approach with an application to the national crime victimization survey data. *Sociological Methods & Research, 31*(4), 514-544.

Mukhopadhyay, K., & Trivedi, P. (1997). Regression Models for Under-Reported Counts. *Unpublished doctoral dissertation, Indiana University, Indiana*.

Neubauer, G., & Djuraš, G. (2008). *A generalized Poisson model for underreporting.* . Paper presented at the Proceedings of the 23rd International Workshop on Statistical Modelling, Utrecht.

Neubauer, G., & Djuraš, G. (2009). *A beta-Poisson model for underreporting.* Paper presented at the Proccedings of the 24rd International Workshop on Statistical Modelling, Ithaca.

Neubauer, G., Duras, G., & Friedl, H. (2010). Models for underreporting: A Bernoulli sampling approach for reported counts.

Özmen, İ. (2000). Quasi likelihood/moment method for generalized and restricted generalized Poisson regression models and its application. *Biometrical Journal: Journal of Mathematical Methods in Biosciences, 42*(3), 303-314.

Pararai, M., Famoye, F., & Lee, C. (2006). Generalized poisson regression model for underreported counts. *Advances and Applications in Statistics, 6*(3), 305 - 322.

Pararai, M., Famoye, F., & Lee, C. (2010). Generalized poisson-poisson mixture model for misreported counts with an application to smoking data. *Journal of Data Science, 8*(4), 607-617.

Ridout, M. S., & Besbeas, P. (2004). An empirical model for underdispersed count data. *Statistical Modelling, 4*(1), 77-89.

Rose, N. L. (1990). Profitability and product quality: Economic determinants of airline safety performance. *Journal of Political Economy, 98*(5, Part 1), 944-964.

Sellers, K. F. (2011). Introducing a Model to Determine True Counts via the Conway-Maxwell-Poisson Distribution. *IWSM 2011*, 548.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., & Boatwright, P. (2005). A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 54*(1), 127-142.

Smita, S. (2009). Statistics and simulation. Retrieved from https://www.moresteam.com/whitepapers/download/stats-and-sim.pdf

Syndromes, S. P. T. f. A. A. C. G. J. J. o. A. I. D. (2008). Project eban: An hiv/std intervention for african american couples. *49*(Suppl 1), S15.

Wang, W., & Famoye, F. (1997). Modeling household fertility decisions with generalized Poisson regression. *Journal of Population Economics, 10*(3), 273-283.

Winkelmann, R. (1995). Duration dependence and dispersion in count-data models. *Journal of Business & Economic Statistics, 13*(4), 467-474.

Winkelmann, R. (1996). Markov chain Monte Carlo analysis of underreported count data with an application to worker absenteeism. *Empirical Economics, 21*(4), 575-587.