Novel Application of a Weighted Zero-Inflated Negative Binomial Model in Modeling Count Data from a Complex Survey

Lin Dai and Mulugeta Gebregziabher*

*corresponding author

Department of Public Health Sciences, Medical University of South Carolina,135 cannon St, MSC 835, Charleston, SC 29425, Tel: 843-876-1112, email: gebregz@musc.edu;

Abstract

We demonstrate a novel application of a weighted zero-inflated negative binomial model to quantify regional variation in HIV-AIDS prevalence in sub-Saharan African countries. We use data from latest round of the Demographic and Health survey (DHS) conducted in three countries (Ethiopia-2011, Kenya-2009 and Rwanda-2010). The outcome is an aggregate count of HIV cases in each census enumeration area (CEA) from the DHS of the three sub-Saharan African countries. Data are characterized by excess zeros and heterogeneity due to clustering. We compare several scale-weighting approaches to account for the complex survey design and clustering in a zero inflated negative binomial (ZINB) model. Finally, we provide marginalized rate ratio (RR) estimates from the best ZINB model.

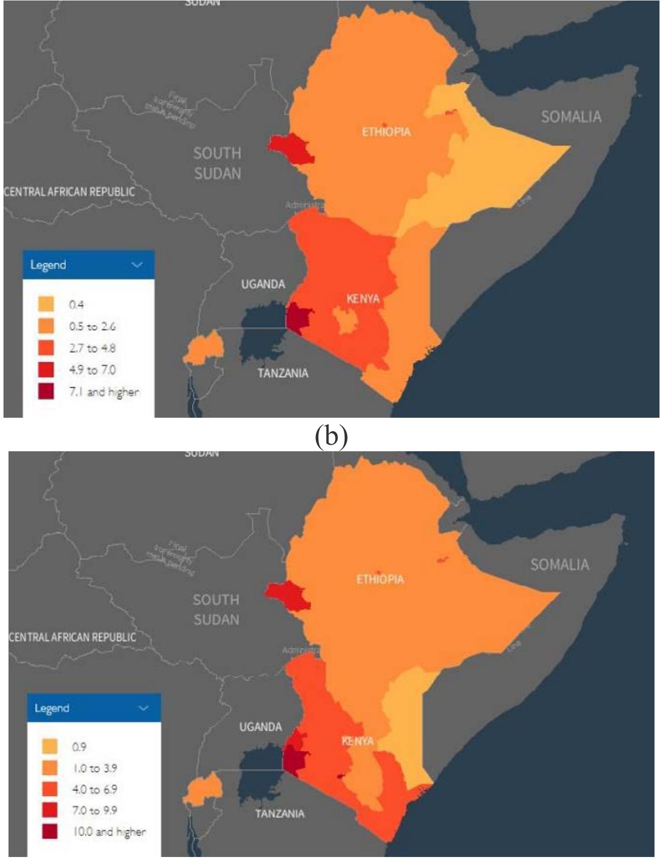
Key words: HIV multi-country survey data, negative binomial, regional variation, zero-inflation

Word Count: Abstract 114; Text 3439; Tables 2; Figures – 4; References 18

1 Introduction

Sub-Saharan Africa has the largest AIDS epidemic in the world. There have been 24.7 million persons infected with HIV in sub-Saharan Africa, which constitutes about 4.7% of the adult population in this region and accounts for around 70% of the people living with HIV worldwide (AVERT 2014). HIV prevalence in sub-Saharan Africa varies regionally. In Kenya (2010) and Ethiopia (2011), HIV prevalence was 6.4% and 1.5%, respectively [Demographic and Health survey (DHS)]; in Rwanda (2010), the prevalence was 3.0% (CSA 2012, Macro 2010, NISR 2012). As depicted in Figure 1a, gender specific HIV prevalence varies not only by country, but also by provinces within each country. For example, in Ethiopia and Kenya, the HIV prevalence among both genders increases from east to west. It is of high importance to policy makers and regional health administrators to understand the characteristics that have influence in regional HIV prevalence and consequently there is an uptake of research towards this goal. However, most of the complex survey studies on HIV prevalence in sub-Saharan African countries are country-specific. Few studies have used multi-country data to assess the issue of regional variation in sub-Saharan Africa. The statistical approaches used in these studies are limited to less advanced statistical methods. The main goal of this study is therefore to show a novel application of a weighted zero-inflated negative binomial model (ZINB) to examine factors (demographic, socio-economic, behavioral, and HIV knowledge) associated with regional variation in HIV-AIDS prevalence in sub-Saharan Africa.

Figure 1a. HIV prevalence in Ethiopia, Kenya and Rwanda ((a) Male (b) Female)


(a)

(b)

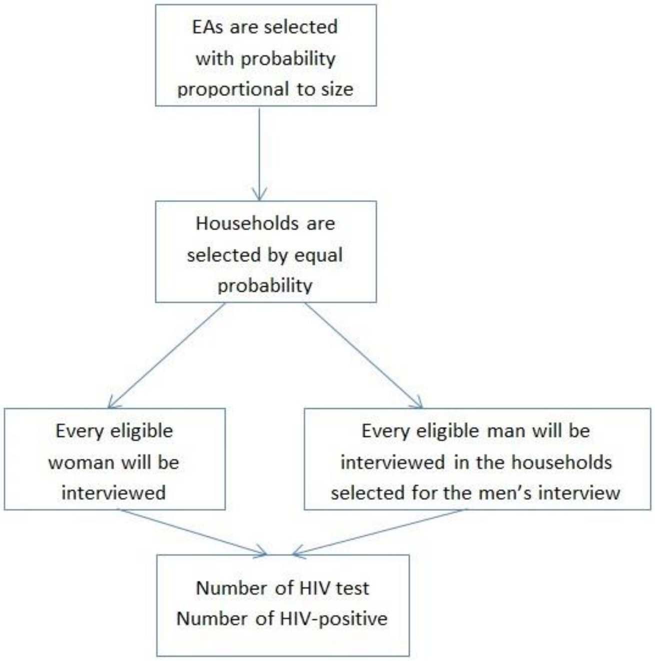
Legend description: prevalence increases with increasing intensity of darkness in color

2 Data and Study Design

Use of a weighted zero-inflated negative binomial model (ZINB) to examine factors associated with regional variation in HIV-AIDS is novel in that, in addition to zero-inflation, the model also accounts for both the complex sampling design nature of the data (two-stage cluster sampling design) and for the clustering of observed count responses by census enumeration area (CEA) and country. Data used in this study are count data from the latest round of DHS conducted from 2008 to 2011in three countries (Ethiopia-2011, Kenya-2009 and Rwanda-2010). The primary outcome is defined as an aggregated count of HIV positive people in each CEA standardized by CEA specific population size as an offset. We compare several scale-weighting approaches to account for the complex survey design and clustering in a zero inflated negative binomial (ZINB) model. We also provide marginalized rate ratio (RR) estimates from the best ZINB model that provide measure of the overall association between HIV prevalence and covariates.

The survey data from each country were obtained via household-based surveys which used a two stage sample design (Figure 2). At the first stage, a sample of CEAs was selected with probability proportional to size, and at the second stage, households were selected by equal probability in the selected CEAs. In each selected households, all women of reproductive age (15–49) were eligible and considered for an individual interview (ICF-International 2012b). In every second or third selected households, eligible men were included for an individual interview. In these selected households, all eligible respondents were asked to give a few drops of blood to be tested in a laboratory for HIV (ICF-International 2012a). The HIV test results of those eligible and who consented were linked to the interview information.

Figure 2 Two stage sample design of DHS


EAs are selected with probability proportional to size
Households are selected by equal probability
Every eligible woman will be interviewed
Every eligible man will be interviewed in the households selected for the men's interview
Number of HIV test
Number of HIV-positive

Variables assessed in the DHS included age, sex, education, and the relationship of the subject to the head of the household among other characteristics. Since the eligibility age ranges for males and females are different, to balance the gender proportion in each CEA, the data used for analysis included eligible participants aged 15 to 49 years who had the HIV test. As shown in Figure 1b, both the outcome variable and

the risk factors are aggregated for each CEA to generate cluster-level information. This allowed us to get more stable values of the variables that were less affected by measurement error (Guthrie, Sheppard, and Wakefield 2002). The primary outcome is defined as an aggregated count of HIV positive people in each CEA. Except for country and location of residence (whether the cluster is urban or rural); the cluster-level variables are derived as the weighted proportion of individuals who have specific characteristics in the cluster. Key variables used in analyses are as follows:
Gender: summarized as percentage of males in each CEA.
Age: percentage of population in each age category in each CEA.
Marital status: percentage of single, married, and divorced members in a cluster.
HIV knowledge: percentage in each category answering correctly to a standard battery of knowledge questions: very low AIDS knowledge (0–20% correct answers), low AIDS knowledge (20–40% correct answers), medium AIDS knowledge (40–60% correct answers), and high AIDS knowledge (60–100% correct answers).
STI symptom prevalence: percentage of people with any self-reported signs or symptoms of sexually transmitted infections, such as genital discharge or genital ulcer.
Multiple sexual partners: percentage within categories relating to the number of self-reported sexual partners in the past 12-months: (1) none, (2) one, (3) two or more.
Media usage: percentage of people who reported using any of the three media (television, radio, or newspaper) more than 1 hour per week

3 Statistical Models and Inference

3.1 Pseudo maximum likelihood for multilevel models

We consider a two stage sampling scenario. At the initial stage, cluster $j = 1,2,...,m$ is sampled with probability π_j , and at the subsequent stage, unit i is sampled with conditional probability $\pi_{i|j}$ given that cluster j was sampled in the first stage. In level two, the sampling weight for cluster j is $w_j = 1/\pi_j$. In level one, the sampling weight for unit i in cluster j is defined as $w_{i|j} = 1/\pi_{i|j}$, and the probability of selection for each unit is computed as $\pi_{ij} = \pi_j \pi_{i|j}$.

For two stage sampling, for individual $i = 1,2,...,n_j$ in cluster $j = 1,2,...,m$, let y_{ij} be the observed variable. Let x_{ij} be the individual level covariates and x_j be the cluster level covariates. Let $f\left(y_{ij} \mid x_{ij}, \eta_j, \theta_1\right)$ be the density function of y_{ij} and $\phi\left(\eta_j \mid x_j, \theta_2\right)$ be the density function of η_j . θ_1 and θ_2 are the parameters to be estimated. Let w_j be the sampling weights for cluster j and $w_{i|j}$ be the conditional sampling weight for individual i in clusters, and s_j be a scale factor. The multilevel pseudo maximum likelihood estimates are the parameters that maximize the weighted pseudo-likelihood (Asparouhov 2006). The corresponding multilevel pseudo maximum likelihood (MPML) can be given by,

$$l(\theta_1,\theta_2) = \prod_j \left(\int \left(\prod_i f\left(y_{ij} \mid x_{ij}, \eta_j, \theta_1\right)^{w_{i|j} s_j} \right) \phi\left(\eta_j \mid x_j, \theta_2\right) d\eta_j \right)^{wj}$$

3.2 Weight scaling method

With the aim of reducing bias in parameter estimates, different scaling methods for the MPML above have been proposed (Stapleton 2002, Pfeffermann et al. 1998). One of the most common ways of scaling weights is to multiply weights by a scale factor so that the sum of the weights is equal to some characteristic of the cluster sample (Potthoff, Woodbury, and Manton 1992). The two most common scaling methods are as follows:

Method A: the sum of the weights is equal to the effective sample size $\frac{\left(\sum_i w_{i|j}\right)^2}{\sum_i (w_{i|j})^2}$ (Longford 1996), and the scale factor for weight of the units in cluster j becomes

$$s_j = \frac{\sum_i w_{i|j}}{\sum_i (w_{i|j})^2}$$

Method B: the sum of the weights is equal to the actual cluster size n_j (Pfeffermann et al. 1998), and the scale factor weight of the units in cluster j becomes

$$s_j = \frac{n_j}{\sum_i w_{i|j}}$$

In this study, we consider and compare four different approaches to incorporating weights in MPML that range from the simplest unscaled raw weights to advanced scale adjusted weights. These scaling methods are described below:

Method 0. Unweighted analysis
Method 1. Raw weight ($w_{i|j}$) is used with $s_j = 1$.

Method 1A. Rescaled weight method A is used with $s_j = \frac{n_j}{\sum_i w_{i|j}}$.

Method 1B. Rescaled weight method B is used with $s_j = \frac{\sum_i w_{i|j}}{\sum_i (w_{i|j})^2}$.

3.3 MPML for zero inflated count data

In a two level dataset, a generalized linear regression model with random effect can be formulated as

$$E\left(y_{ij} \mid x_{ij}, \eta_j, \theta_1\right) = \mu_{ij} = g^{-1}(\beta_0 + \beta_1 x_{ij} + \eta_j)$$

where y_{ij} is the dependent variable, x_{ij} are the individual level covariates, $\theta_1 = (\beta_0, \beta_1)$ are the parameters to be estimated, $\eta_j \sim N(0, \sigma_1^2)$ is the cluster level random effect, and g is a monotone link function (described below).

Typically, Poisson regression is used to model count data where observations are assumed to be independent and the number of cases has variance equal to the mean for each level of the covariates. However, in a complex survey situation, often either the independence or the equal mean and variance assumption is violated, mostly leading to overdispersion. Thus, we consider a negative binomial (NB) model that handles the problem of overdispersion (Moghimbeigi et al. 2008) via inclusion of an extra parameter. To accommodate the case where overdispersion may not be sufficiently modeled via the extra parameter in the NB model, we consider including random effects into the NB model.

Another challenge with modeling count data is the issue of excess zeroes. When count data that are distributed as NB have point mass at zero, the zero-inflated negative binomial (ZINB) model handles the problem of excess zeros. The ZINB model is a mixture of the NB model for the count part (Y_{ij}) and a logit model for the excess zeros. Assuming an observation has probability p_{ij} to be zero and probability $1-p_{ij}$ to follow a negative binomial distribution, the ZINB distribution is given by (Lambert 1992),

$$f(y_{ij}=0)=p_{ij}+(1-p_{ij})(1+\frac{\lambda_{ij}}{r})^{-r}$$

$$f(y_{ij}=y)=(1-p_{ij})\frac{\Gamma(y+r)}{y!\Gamma(r)}(1+\frac{\lambda_{ij}}{r})^{-r}(1+\frac{r}{\lambda_{ij}})^{-y},\ y=1,2,...$$

Since zero inflated models correspond to a mixture of a subpopulation generated from a negative binomial distribution and a subpopulation that provides the excess zeros, it is difficult to use the estimated parameters to make inference on the marginal mean of the sampled population. The marginalized zero-inflated model has been developed to directly model the marginal means of mixtures of two discrete distributions to make the inference straightforward. The marginalized zero-inflated model of a two level dataset can be formulated as a two part model as follows (Long et al. 2014):

$$\log\{p_{ij}/(1-p_{ij})\}=\alpha_0+\alpha_1 X_{ij}+\eta_{1j}$$

$$\log(\lambda_{ij})=\log(m_{ij})+\log(1+\exp(\alpha_0+\alpha_1 X_{ij}+\eta_{1j}))$$

where

$$m_{ij}=\exp(\log(N_{ij})+\beta_0+\beta_1 X_{ij}+\eta_{ij})$$

is the overall mean, and $(\eta_{1j,}\eta_{2j})$ are level two random effects, and N_{ij} is included as an offset term.

3.4 Variance Estimation

If $f(\theta)$ is the likelihood for the optimization, then the asymptotic covariance matrix of the maximum likelihood estimator is given by,

$$\text{cov}\left(\hat{\theta}\right)=H\left(\hat{\theta}\right)^{-1}\left(\sum_i g_i(\hat{\theta})g_i(\hat{\theta})'\right)H\left(\hat{\theta}\right)^{-1}$$

where $H\left(\hat{\theta}\right)$ is the second derivative matrix of $f(\theta)$ and $g_i(\hat{\theta})$ is the first derivative of $f(\theta)$ for the i th subject (Self and Liang 1987).

In standard maximum likelihood estimation, the covariance estimator is obtained by inverting the information matrix, $\mathrm{H}(\theta)$. Obtaining the sandwich estimator is also straightforward. However, when using pseudo-likelihood estimation, the sandwich estimator does not collapse because the pseudo-likelihood does not represent the distribution of responses (Rabe-Hesketh and Skrondal 2008). Both Proc GLIMMIX and NLMIXED procedures in SAS are used for estimating the sandwich variance using the EMPIRICAL option in PROC NLMIXED and the EMPIRICAL=CLASSICAL option in PROC GLIMMIX (SAS 2004).

4 Simulation Study

4.1 Simulation

To generate data under the zero inflated negative binomial distribution, we used the conditional mean $E(Y_{ij})=\mu_{ij}=g^{-1}(\beta_0+\beta_1 x_{ij}+\beta_2 x_j+\eta_{1j})$ with $\beta_0=\beta_1=\beta_2=1$ and a zero-inflated model given by $\text{logit}(\alpha_0+\alpha_1 x_{ij}+\alpha_2 x_j+\eta_{0j})$. We also assumed the following $\alpha_1 = \alpha_2 = 0.5$; $\eta_j = c(\eta_{1j}, \eta_{2j})$, which is the cluster level random intercept and

$\eta_j \sim N\left((0,0)^T, \begin{pmatrix}\sigma_1^2 & 0 \\ 0 & \sigma_2^2\end{pmatrix}\right)$ with $\sigma_1^2 = 0.25$ and $\sigma_2^2 = 0.09$. We use three different values for α_0 to get three different zero inflation proportions. For α_0 = -2, -1, -0.2 the approximate proportion of zeros are 15%, 30% and 45%, respectively. The joint probability function for $Y_{ij} = y_{ij}$ is given by,

$$P(Y_{ij} = y_{ij}) = \begin{cases} \text{logit}(\alpha_0 + \alpha_j x_j + \eta_{0j}) + (1 - \text{logit}(\alpha_0 + \alpha_j x_j + \eta_{0j}))k(0) & y_{ij} = 0 \\ (1 - \text{logit}(\alpha_0 + \alpha_j x_j + \eta_{0j}))k(y_{ij}) \quad y_{ij} > 0 \end{cases},$$

where $k(.)$ is either a Poisson distribution or negative binomial distribution.

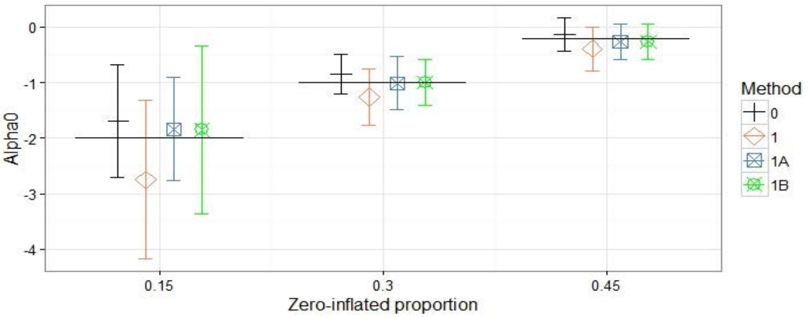
The selection probability of each cluster j is defined by $\pi_j = 1/4$ if $|\eta_j| \le 1$, and $\pi_j = 3/4$ if $|\eta_j| > 1$, and the selection probability of each unit ij in each cluster j is defined by $\pi_{i|j} = 1/2$ if $y_{ij} > 0$, and $\pi_{i|j} = 3/4$ if $y_{ij} = 0$.

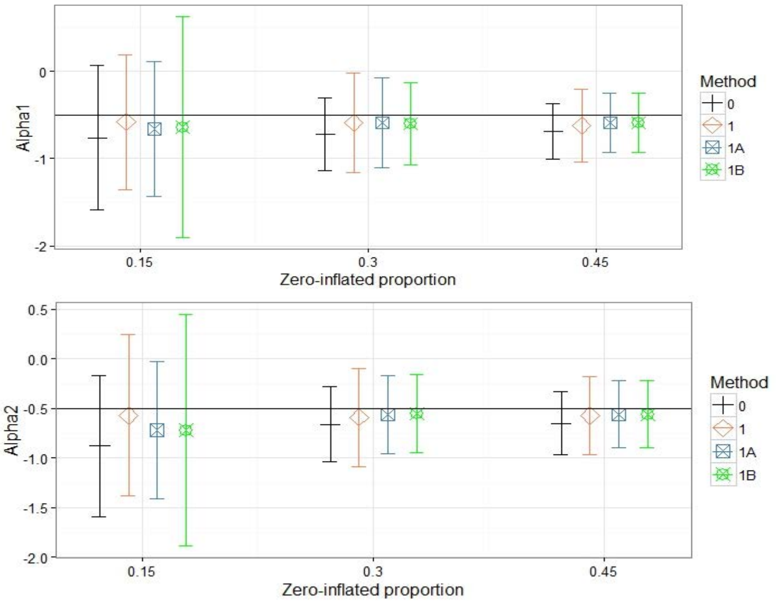
The NLMIXED procedure in SAS 9.4 is used to perform the pseudo-likelihood estimation using ZINB model. Each of the analysis is replicated 100 times, and in all cases, we generate datasets which contain 100 cluster units. We compare the performance of the different scaling methods in terms of bias, asymptotic standard error.

4.2 Simulation results

The results of the ZINB model with different zero inflation proportions are given in Figure 3 and Figure 4. For coefficients in the zero part, using scaled sampling weights (Method 1A and 1B) gives the least biased estimates. For coefficients (β_1 and β_2) in the count part, when the zero inflation proportion is 15%, the least biased method for estimating parameters is the raw weighted model (Method 1), and when the zero inflation proportion is 30% to 45%, the least biased method for estimating parameters is the scaled weighted model (Method 1A and 1B). As expected, with increasing zero inflation, all the models give better estimations for the coefficients in the zero part.

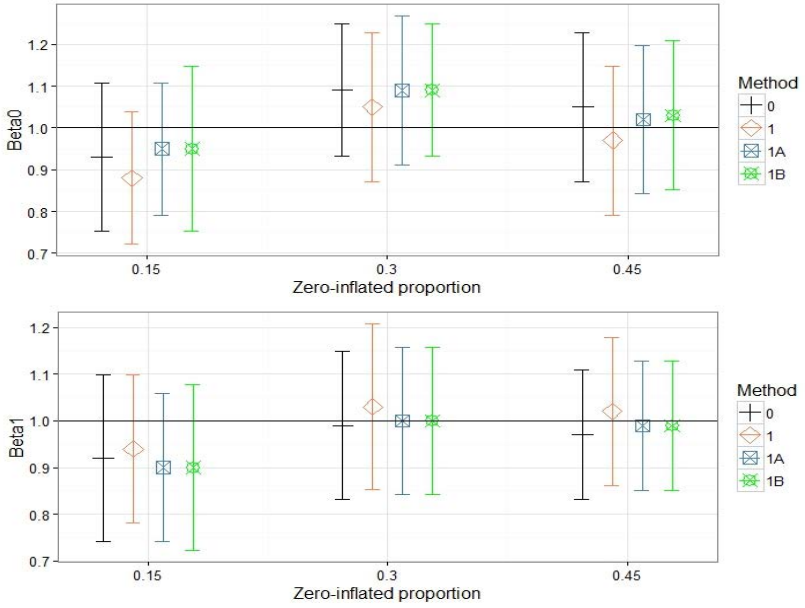
Figure 2 Parameter estimate for ZINB model (zero-inflated part)

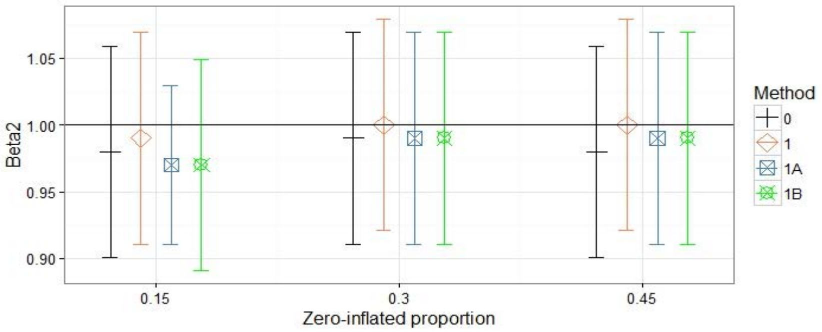




For the coefficients (β_0 , β_1 and β_2) in the count part, the estimates from Methods 1A and 1B are similar. However, the zero inflation proportion seems to affect the estimates for all four weighting methods. The most biased scenario is the one that has the smallest zero proportion. When the zero proportion is 15% all the methods give biased estimates for the intercepts in the zero part. This may be due to the fact that these models account for these through the overdispersion parameter of the NB model. For coefficients β_0 and β_1, the least biased estimates result when using scaled weights (Method 1A and 1B). However, for coefficient β_2 using scaled weights (Method 1A and 1B) gives the least biased estimates. Overall, while there does not appear to be one method that applies to all situations, the ZINB approach seems more robust.

Figure 3 Parameter estimate for ZINB model (count part)





5 DHS Data Analysis Results

Table 1 shows the parameter estimates using (unweighted) maximum likelihood and pseudo-maximum-likelihood. Although the models resulted in different estimates of the parameters, most of them identified country, gender, rural residence, proportion with STI, number of partners, and marital status to be significantly associated with HIV prevalence (p<0.05). Compared to the weighted models, the unweighted model was more likely to generate estimates with greater standard error, which resulted in the weighted methods giving more significant results compared to the unweighted method.

Table 1 Parameter estimates of zero-inflated negative binomial (ZINB) models; DHS survey results for Ethiopia, Kenya and Rwanda; 2008–2011.

		Non weight			Raw weight			Scale Method A			Scale Method B		
Parameters		Est	Std	P	Est	Ste	P	Est	Std	P	Est	Std	P
Count part													
Country	Kenya	1.82	0.72	0.02	1.73	0.16	<.01	1.51	1.27	0.25	1.55	0.60	0.02
	Rwanda	1.16	0.44	0.01	1.07	0.17	<.01	1.19	0.70	0.10	1.13	0.36	<.01
gender	male	-0.70	0.45	0.13	-1.27	0.05	<.01	-1.22	0.67	0.08	-1.24	0.36	<.01
media use	Yes	0.58	0.57	0.31	1.12	0.07	<.01	1.52	0.90	0.11	1.39	0.51	<.01
STI burden	Ethiopia	6.92	2.23	0.01	6.94	0.24	<.01	7.12	4.09	0.09	7.22	1.62	<.01
	Kenya	-6.75	2.53	0.01	-7.06	0.28	<.01	-7.54	4.58	0.11	-7.52	1.88	0.00
	Rwanda	-3.85	2.46	0.13	-4.29	0.26	<.01	-4.57	4.30	0.30	-4.66	1.87	0.02
HIV Knowledge	20-40%	-1.70	1.52	0.27	-1.19	0.19	<.01	-1.88	2.41	0.44	-1.05	1.31	0.43
	40-60%	-1.08	1.37	0.44	-0.44	0.18	0.02	-1.37	2.20	0.54	-0.35	1.22	0.78
	>=60%	-0.04	1.33	0.98	0.17	0.18	0.33	-0.87	2.19	0.69	0.04	1.21	0.97
Partners	1	1.73	0.57	0.01	2.07	0.06	<.01	2.00	0.91	0.04	2.13	0.47	<.01
	2+	2.07	0.97	0.04	1.59	0.12	<.01	1.43	1.61	0.38	1.44	0.88	0.11
Marital Status	Married	0.84	0.80	0.31	1.52	0.09	<.01	0.81	1.33	0.55	1.01	0.65	0.13
	Divorced	5.41	1.36	0.00	2.05	0.15	<.01	0.88	2.40	0.72	1.34	1.04	0.21
Age	20-24	-0.47	0.60	0.45	-0.94	0.07	<.01	-0.42	0.89	0.64	-0.73	0.50	0.16
	25-29	-1.19	0.65	0.08	-1.28	0.08	<.01	-0.70	1.00	0.49	-1.00	0.59	0.10
	30-34	-0.66	0.70	0.36	-1.16	0.08	<.01	-0.30	1.05	0.77	-0.73	0.61	0.24
	35-39	-0.45	0.78	0.57	-1.18	0.09	<.01	-0.92	1.21	0.45	-1.29	0.68	0.07
	40-44	-1.26	0.85	0.15	-1.85	0.10	<.01	-1.80	1.27	0.17	-2.05	0.76	0.01
	45+	-2.63	0.87	0.01	-2.63	0.10	<.01	-2.81	1.33	0.05	-2.91	0.74	<.01
Education	Primary	0.47	0.48	0.34	0.55	0.06	<.01	-0.14	0.73	0.85	0.25	0.42	0.56

	Secondary	0.47	0.53	0.39	0.27	0.06	<.01	-0.40	0.81	0.62	-0.17	0.46	0.72
	More than secondary	-0.23	0.64	0.72	-0.14	0.07	0.06	-1.05	1.05	0.33	-0.58	0.55	0.30
Residence	Ethiopia	-0.87	0.20	0.00	-0.22	0.02	<.01	-0.39	0.32	0.25	-0.34	0.14	0.03
	Kenya	0.75	0.26	0.01	0.20	0.03	<.01	0.34	0.44	0.44	0.24	0.20	0.23
	Rwanda	0.10	0.25	0.69	-0.43	0.03	<.01	-0.56	0.37	0.15	-0.46	0.20	0.03
Zero part													
Country	Kenya	6.30	13.25	0.64	-0.93	0.04	<.01	-0.87	0.44	0.06	-0.91	0.27	<.01
	Rwanda	-1.81	46.89	0.97	-1.63	0.06	<.01	-1.38	0.44	0.00	-1.46	0.49	0.01
STI burden		-9.33	8.65	0.29	-6.05	0.39	<.01	-6.65	3.92	0.10	-6.62	2.75	0.02
Partners	1	2.55	3.20	0.43	2.90	0.20	<.01	2.53	2.21	0.26	2.82	1.39	0.05
	2+	-7.50	8.99	0.41	-3.13	0.40	<.01	-2.67	4.23	0.53	-2.70	2.72	0.33
Marital Status	Married	-4.56	3.23	0.17	0.02	0.18	0.90	0.07	1.98	0.97	-0.09	1.22	0.94
	Divorced	-10.24	8.46	0.24	-8.13	0.30	<.01	-8.88	3.44	0.02	-8.55	2.06	<.01

Note: Est: Estimate Std: Standard error P: P-value Reference group for each variables: country: Ethiopia, gender: female, HIV knowledge: answer less than 20%, age: 15–19, education: no education, STI: no STI in each CEA, marital status: singles in each CEA, residence: urban in each CEA, number of partners: zero partners.

Table 2 shows the rate ratio (RR) and 95% CI estimates from the weighted ZINB and marginal ZINB (mZINB) models using method A. The RR of mZINB can be interpreted as a regular RR estimate from a NB model. Higher proportion of HIV knowledge, higher number of sexual partners, higher proportion of people with STI, and rurality of residence were all associated with HIV prevalence. Compared to those CEAs with a higher proportion of people with no sexual partners, CEAs with a higher proportion of people with one or more partners were associated with higher prevalence of HIV (RR=3.58 for two or more and RR=3.09 for one or more). Compared to CEAs with higher proportion of people with less than 20% HIV knowledge, CEAs with higher proportion of people with HIV knowledge more than 20% were associated with lower prevalence of HIV (RR=0.05 for 20%-40% RR=0.10 for 40-60% and RR=0.13 for >60%). CEAs with a higher proportion of STI (p<0.05), and that are rural (p<0.05) are associated with HIV prevalence differentially by country. In Ethiopia, Kenya and Rwanda, CEAs with higher proportion of people with STI compared to those without STI are 1031.12, 1.72 and 55.26 times more likely to have higher prevalence of HIV, respectively. Rurality of the census enumeration area was also associated with lower rate of HIV prevalence in Rwanda (RR=0.46), while in Ethiopia and Kenya there was no significant difference between the people living in urban and rural areas.

Table 2 also shows odds ratio estimates from the ZINB and mZINB models in zero part. The two models generated similar values. Each estimate compares the odds of being an excess zero to the odds of not being an excess zero in each CEA as a function of selected covariates. Clusters with an excess of negative tests (zeros) were more likely to have a lower number of people with two or more partners and a lower proportion of people with STI.

Table 2 Parameter estimates of zero-inflated negative binomial (ZINB) and marginal ZINB Models for HIV prevalence

		ZINB			Marginal ZINB		
		RR	95% CI		RR	95% CI	
Count part							
media use		4.55	0.7	29.38	2.68	0.95	7.5
HIV Knowledge	20-40%	0.15	0	22.44	0.05	0	0.7
	40-60%	0.25	0	24.01	0.1	0.01	1.38
	>=60%	0.42	0	38.64	0.13	0.01	1.66
Partners	1	7.38	1.11	48.89	3.09	1.11	8.6
	2+	4.16	0.15	115.7	3.58	0.57	22.59
Education	Primary	0.87	0.19	3.95	1.2	0.53	2.71
	Secondary	0.67	0.12	3.58	0.84	0.34	2.06
	More than secondary	0.35	0.04	3.09	0.42	0.14	1.22
STI burden	Ethiopia	911.87	26.33	31583.81	1031.12	33.65	31602.77
	Kenya	0.65	0.01	43.12	1.72	0.2	14.74
	Rwanda	12.76	0.6	273.66	55.26	7.79	391.86
Residence	Ethiopia	0.68	0.35	1.33	0.78	0.59	1.03
	Kenya	0.96	0.5	1.84	0.91	0.67	1.23
	Rwanda	0.39	0.25	0.62	0.46	0.34	0.63
Zero part		OR			OR		
Country	Kenya	0.42	0.17	1.03	0.41	0.25	0.68
	Rwanda	0.25	0.1	0.63	0.31	0.15	0.63
STI burden		0	0	4.27	0.03	0	5.44
Partners	1	12.61	0.13	1223.9	22.17	1.61	305.39
	2+	0.07	0	437.77	0.01	0	6.47

Note: Est: Estimate Std: Standeard error P: P-value Reference group for each variables: country: Ethiopia, HIV knowledge: answer less than 20%, age: 15–19, education: no education, STI: no STI in each CEA, marital status: singles in each CEA, residence: urban in each CEA, number of partners: zero partners.

6 Conclusion and Discussion

In this study, we demonstrated a novel application of a scale weighted ZINB to analyze count data from a survey of three countries. This expands current single country level approaches to include incorporation of aggregate data to examine factors associated with regional variation in HIV-AIDS prevalence in sub-Saharan Africa. We also made comparisons among several weight scaling methods for incorporating design weights in a multilevel pseudo likelihood via simulation studies. Our results show the utility of multilevel pseudo likelihood estimation in complex survey data when design weights are properly accounted for by considering appropriate scaling methods. We also estimated RRs of the mZINB model to report estimates that have the same interpretation as regular RRs from NB models. Based on the analysis results, we conclude that scale weighted ZINB model is effective and robust for the analysis of aggregated count data with extra heterogeneity due to clustering. Finally, we would like to mention that further simulation studies might be needed to fully understand the operational characteristic of these models. Applications in other areas of biomedical research in which count responses are measured in clusters using a complex survey design might also be helpful to get further insights about the applications of these models.

References

Asparouhov, Tihomir. 2006. "General multi-level modeling with sampling weights." *Communications in Statistics—Theory and Methods* no. 35 (3):439-460.

AVERT. *HIV and AIDS in Sub-saharan Africa.* 2014 [cited 14 July 2016.

CSA, ICF-International. 2012. "Ethiopia demographic and health survey 2011." *Addis Ababa, Ethiopia and Calverton, Maryland, USA: Central Statistical Agency and ICF International*.

Guthrie, Katherine A, Lianne Sheppard, and Jon Wakefield. 2002. "A hierarchical aggregate data model with spatially correlated disease rates." *Biometrics* no. 58 (4):898-905.

ICF-International. 2012a. "Biomarker Field Manual." *Calverton, Maryland, USA . Demographic and Health Surveys*.

ICF-International. 2012b. Demographic and Health Survey Sampling and Household Listing Manual. . In *MEASURE DHS, Calverton, Maryland, U.S.A.: ICF International* ICF International.

Lambert, Diane. 1992. "Zero-inflated Poisson regression, with an application to defects in manufacturing." *Technometrics* no. 34 (1):1-14.

Long, D Leann, John S Preisser, Amy H Herring, and Carol E Golin. 2014. "A marginalized zero-inflated Poisson regression model with overall exposure effects." *Statistics in medicine* no. 33 (29):5151-5165.

Longford, Nicholas T. 1996. "Model-based variance estimation if surveys with stratified clustered design." *Australian Journal of Statistics* no. 38 (3):333-352.

Macro, ICF-International. 2010. *Kenya Demographic and Health Survey 2008-09*: KNBS.

Moghimbeigi, Abbas, Mohammed Reza Eshraghian, Kazem Mohammad, and Brian Mcardle. 2008. "Multilevel zero-inflated negative binomial regression modeling for over-dispersed count data with extra zeros." *Journal of Applied Statistics* no. 35 (10):1193-1202.

NISR, MOH, and ICF-International. 2012. "Rwanda Demographic and Health Survey 2010." *Calverton, Maryland, USA NISR, MOH, and ICF International.* .

Pfeffermann, Danny, Chris J Skinner, David J Holmes, Harvey Goldstein, and Jon Rasbash. 1998. "Weighting for unequal selection probabilities in multilevel models." *Journal of the Royal Statistical Society: series B (statistical methodology)* no. 60 (1):23-40.

Potthoff, Richard F, Max A Woodbury, and Kenneth G Manton. 1992. "“Equivalent sample size” and “equivalent degrees of freedom” refinements for inference using survey weights under superpopulation models." *Journal of the American Statistical Association* no. 87 (418):383-396.

Rabe-Hesketh, Sophia, and Anders Skrondal. 2008. *Multilevel and longitudinal modeling using Stata*: STATA press.

SAS, II. 2004. "SAS/STAT® 9.1 user's guide." *SAS Institute Ine Cary, NC.*

Self, Steven G, and Kung-Yee Liang. 1987. "Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions." *Journal of the American Statistical Association* no. 82 (398):605-610.

Stapleton, Laura M. 2002. "The incorporation of sample weights into multilevel structural equation models." *Structural Equation Modeling* no. 9 (4):475-502.