JRHS
Journal of Research in Health Sciences
journal homepage: www.umsha.ac.ir/jrhs

Original Article

Social Hidden Groups Size Analyzing: Application of Count Regression Models for Excess Zeros

Maryam Jalali (MSc)[a], Ali Nikfarjam (MD)[b], Ali Akbar Haghdoost (PhD)[a], Nadere Memaryan (MD)[b], Termeh Tarjoman (MD)[b], and Mohammad Reza Baneshi (PhD)[c*]

[a] *Regional Knowledge Hub for HIV/AIDS Surveillance, Institiue for Futures Studies in Health, Kerman University of Medical Sciences, Kerman, Iran*
[b] *Department of Addiction, Deputy of Mental and Social Health, Health Deputy of Ministry of Health and Medical Education, Tehran, Iran*
[c] *Research Center for Modeling in Health, Institiue for Futures Studies in Health, Kerman University of Medical Sciences, Kerman, Iran*

ARTICLE INFORMATION

Article history:
***Received**: 30 January 2013*
***Revised**: 27 March 2013*
***Accepted:** 02 May 2013*
***Available online:** 14 May 2013*

Keywords:
Regression
Over-dispersion
Negative Binomial distribution
Alcoholics
Poisson

**** Correspondence***
Mohammad Reza Baneshi (PhD)
***Tel:** +98 913 4423948*
***Fax:** +98 341 3205127*
***E-mail1:** rbaneshi@yahoo.com*
***E-mail2:** m_baneshi@kmu.ac.ir*

ABSTRACT

Background: In the case of sensitive questions such as number of alcoholics known, majority of respondents might give an answer of zero. Poisson regression model (P) is the standard tool to analyze count data. However, P provides poor fit in the case of zero inflated counts, when over-dispersion exists. Therefore, the questions to be addressed are to compare performance of alternative count regression models; and to investigate whether characteristics of respondents affect their responses.

Methods: A total of 700 participants were asked about number of people they know in hidden groups; alcoholics, methadone users, and Female Sex Workers (FSW). Five regression models were fitted to these outcomes: Logistic, P, Negative Binomial (NB), Zero Inflated Poisson (ZIP), and Zero Inflated Negative Binomial (ZINB). Models were compared in terms of Likelihood Ratio Test (LRT), Vuong, AIC and Sum Square of Error (SSE).

Results: Percentages of zero were 35% for number of alcoholics, 50% for methadone users, and 65% for FSWs. ZINB provided the best fit for alcoholics, and NB provided the best fit for other outcomes. In addition, we noticed that young respondents, male and those with low education were more likely to know or reveal sensitive information.

Conclusion: Although P is the first choice for modeling of count data in many cases, it seems because of over-dispersion of zero inflated counts in the case of sensitive questions, other models, specifying NB and ZINB, might have better goodness of fit.

Citation: Jalali M ,Nikfarjam A ,Haghdoost AA ,Memaryan N ,Tarjoman T ,Baneshi MR. Social Hidden Groups Size Analyzing: Application of Count Regression Models for Excess Zeros. J Res Health Sci. 2013;13(2): xxxx .

Introduction

A group such as 'number of alcoholics' (in counties where alcohol is totally banned), or 'number of Female Sex Workers (FSW)' are known as 'hidden' or 'hard to reach' sub-groups. Accessing to the members of these groups is difficult. However, information about their size is important for health planning.

The Network Scale-Up (NSU) method is an approach for size estimation of hidden groups [1]. In the NSU we select a random sample from a general population and ask subjects the number of people they know belonging to specific sub-groups such as the number of drug users in their networks. However, majority of responses would be zero, means that in many cases, most of our subjects know nobody in the sub-groups of interest. This can be due to the fact that the real sizes of such sub-groups are usually small. In addition, such groups have low social acceptability (due to their stigmatized nature). Therefore, small proportions of respondents are willing to reveal such information [1].

In the case of NSU studies, we usually face count data with a spike at zero. There are varieties of methods for analyzing count data. In the easiest case, one can consider count outcome as a continuous variable and apply linear regression. However, due to skewed distribution of responses, the normality assumption does not hold [2, 3]. Instead, one can recode the data in to zero (for those who does not know any one) and one (for those who know at least one member of hidden group) and apply logistic regression. However, this method does not make the most use of data [2, 4, 5].

Poisson regression (P) is a standard method for analyzing count data. One assumption for Poisson distribution is the equality of mean and variance. But this condition may not always meet [2,3]. In many count data, the variance is much greater than the mean. This condition is called over-dispersion [2,4,6]. Applying P when over-dispersion is observed leads to biased parameter estimation, too narrow confidence intervals and *P*-values that are too small [4].

Negative Binomial (NB) regression model is an alternative for P when over-dispersion is observed in the dataset [2-4]. This model estimates the correct SE, taking into account the over-dispersion in the data.

However, one more problem in modeling of such count data is that the number of observed zeros might be higher than that expected from a Poisson model (known as "excess zero") [4]. To tackle this problem, Zero Inflated Poisson (ZIP) model can be implemented [4]. When over-dispersion and excess zero happen together, Zero Inflated Negative Binominal (ZINB) regression model is proposed. As in the ZIP regression, in ZINB regression the zero and non zero counts are modeled in different ways [7].

In this paper we have two inter-related aims: to compare the performance of count regression models, in terms of goodness of fit, and to investigate the characteristics of those who know or reveal stigmatized behaviors of people in their network.

Methods

Design of Study

We have used the data of a national NSU study which aimed to estimate the size of drug users and alcoholics in Iran (confidential data submitted to Iranian Ministry of Health and Medical Education (IMHME)). For this study, getting the permission from IMHME, we have used the data of two provinces (Kerman and Fars). In each province about 75% of data were collected from capital and 25% from one main city. Questionnaires were filled through gender-match face to face interviews. Participants were selected from the downtown and uptown, among pedestrians who walked alone. To get a representative sample, we selected pedestrians with different demographic characteristics. We asked subjects about number of people they knew belonging to specific sub-groups.

Outcome and Independent Variables

To compare performance of modeling approaches at different zero percentages, three outcome variables were selected with low, moderate, and high proportion of zero. Response variables (proportion of zeros) were number known to drink alcohol (34%), number known to receive methadone (50%), and number known to be Female Sex Workers (FSW, 65%). For all three outcomes, we asked respondents to count those even with one episode of act in the last year.

The independent variables were gender (male, female), education (under diploma, diploma, undergraduate degree postgraduate degree), age group (<29, 30-39, >40), marital status (single, ever married), and province (Fars, Kerman).

Modeling approaches and Comparison Criterion

Although logistic regression does not suit count data, at the first step we recoded the data to investigate whether logistic model was a good surrogate for count regression models. To do so we recoded the data in to zero (knowing no body at all) and one (knowing at least one).

Then, four models were fitted based on the observed counts; P, NB, ZIP, and ZINB. In addition, model comparison for nested (P versus NB, and ZIP versus ZINB) and non-nested models (P versus ZIP, and NB versus ZINB) has been done through the LRT and Vuong test (V test) respectively. Applying LRT and Vuong tests, the best model was selected and its results were considered as gold standard. In particular we have focused on estimated SEs, and on significance of the variables in each model. Difference more than 15% between SE in gold model and other models was considered as bias.

For all models, AIC were reported. In addition, Sum Square Error (SSE) was estimated as the summation of square difference between real and estimated values.

Additionally a visual depiction of difference between observed versus predicted probabilities among count models were displayed. In all analyses, *P*-value less than 5% was considered as statistical significance. The analysis has been done in stata software version 11.

Results

Descriptive statistics

In total, 327 (46.7%) and 373 (53.3%) of the respondents were male and female respectively. Nearly 80% of our sample was comprised of subjects with diploma or higher levels of education. The age of about 60% of respondents was between 15-29 years old; while the age of about one fifth was more than 40. In addition, nearly 40% of respondents were single.

The histograms of outcome variables are shown in Figure 1, 2 and 3. Of the 700 respondents, 33.9%, 49.1% and 65.3% reported knowing no alcoholics, methadone users and FSWs respectively. Knowing FSW was the most sensitive question which the percent of zeros in the responses was 65%. In addition, 75% of respondents knew at most one FSW. Means (variances) of these outcomes were 8.06 (279.56), 3.01 (87.38), and 1.77 (30.39) respectively, indicating over-dispersion in all outcomes. The descriptive statistics of responses stratified by gender are shown in Table 1.

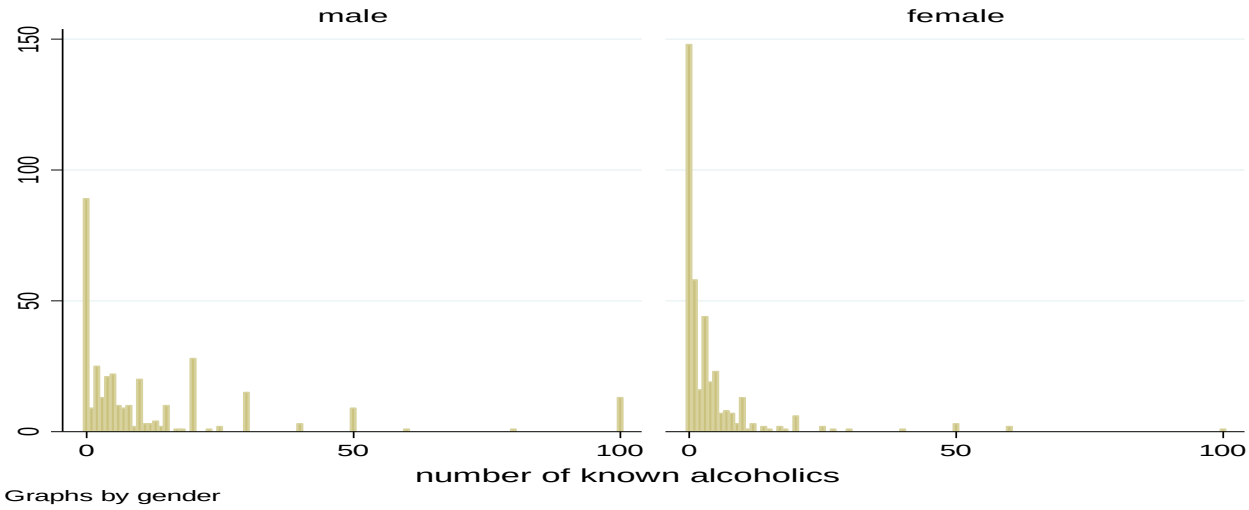

Figure 1: Distribution of number of alcoholics known by respondents, separated by gender

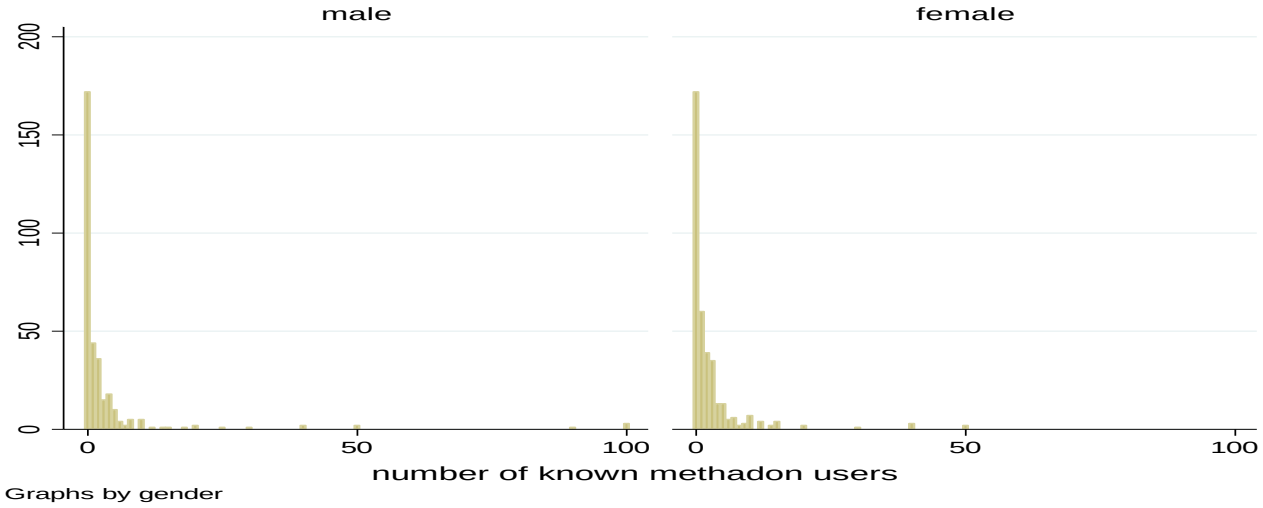

Figure 2: Distribution of number of methadone users known by respondents, separated by gender

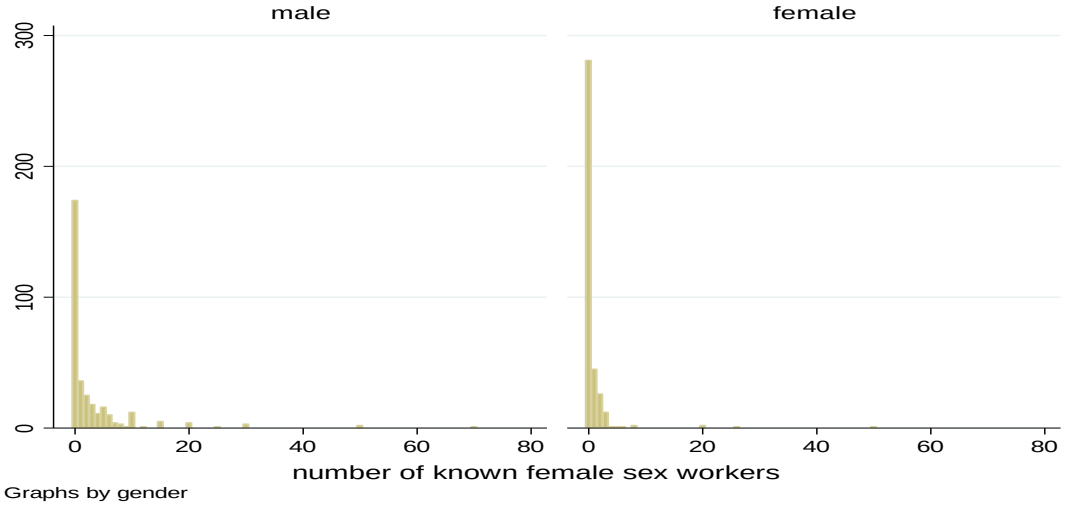

Figure 3: Distribution of number of female sex workers known by respondents, separated by gender

Table 1: Descriptive statistics of responses to sensitive questions stratified by gender

	Alcoholics known		Methadone users known		Female sex workers known	
Variable	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**
Mean	12.66	4.03	3.47	2.06	2.93	0.75
Variance	462.59	85.02	145.70	36.15	49.68	11.36
Minimum	0	0	0	0	0	0
Maximum	100	100	100	50	70	50
1st quartile (Q1)	0	0	0	0	0	0
2nd quartile (Q2)	5	1	0	1	0	0
3rd quartile (Q3)	15	4	2	3	3	0

Factors influence knowing alcoholics

For the alcohol, LRT test suggested that NB was superior to P, indicating existence of over-dispersion (P<0.001). Applying Vuong test, the ZINB model reflected the observed data better than NB (P=0.010) indicating presence of excess zero. However, the AIC and SSE corresponding to ZINB and NB models were comparable (5.55 vs. 5.57), and (0.22 versus 0.26).

According to the Figure 4 it is obvious that the P model dose not predict well the proportion of zeros. This figure shows appropriate fit of ZINB model to the data.

In addition, we have seen that SEs of all variables in P and ZIP models were considerably smaller than ZINB (Table 2). This is mainly due to the fact that P and ZIP models do not consider the over-dispersion parameter. On the other hand, SEs' corresponding to NB and ZINB were fairly similar.

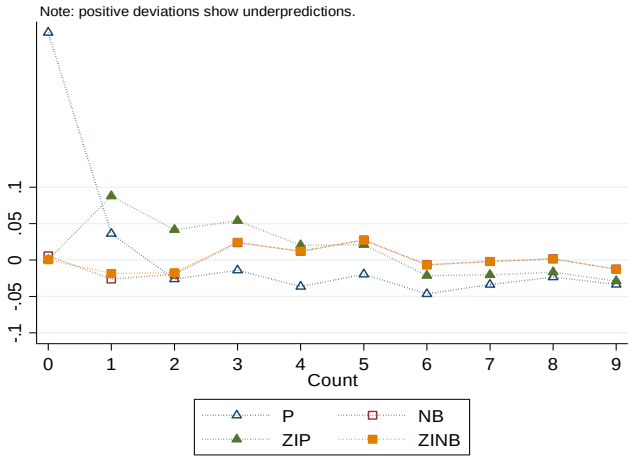

Figure 4: Comparisons among observed versus predicted probabilities among count models for alcohol

Table 2: Influence of demographic characteristics of respondents about number of alcoholics in their network considering zero inflated negative binomial as gold standard; the bold figures mean SEs estimated with bias (>15% difference with the gold model)

	Logistic			Poisson			Negative binomial			Zero inflated Poisson			Zero inflated Negative binomial		
Variables	**0R**	**SE**	***P* value**	**IRR**	**SE**	***P* value**	**IRR**	**SE**	***P* value**	**IRR**	**SE**	***P* value**	**IRR**	**SE**	***P* value**
Province															
Fars		1.00			1.00			1.00			1.00			1.00	
Kerman	0.72	**0.12**	0.053	0.93	**0.03**	0.004	0.77	0.10	0.033	1.06	**0.03**	0.024	0.74	0.09	0.015
Gender															
Male		1.00			1.00			1.00			1.00			1.00	
Female	0.64	**0.11**	0.011	0.34	**0.01**	0.001	0.33	0.04	0.001	0.39	**0.01**	0.001	0.30	0.04	0.001
Marital status															
Single		1.00			1.00			1.00			1.00			1.00	
Ever married	0.82	0.17	0.349	1.00	**0.03**	0.980	1.16	0.19	0.377	1.04	**0.04**	0.321	1.19	0.19	0.283
Age (yr)															
15-29		1.00			1.00			1.00			1.00			1.00	
30-39	0.68	**0.15**	0.079	0.51	**0.02**	0.001	0.50	0.09	0.001	0.59	**0.03**	0.001	0.52	0.09	0.001
≥40	0.35	**0.09**	0.001	0.54	**0.03**	0.001	0.47	**0.10**	0.001	0.83	**0.04**	0.001	0.68	0.16	0.098
Education															
Under diploma		1.00			1.00			1.00			1.00			1.00	
Diploma	0.84	**0.21**	0.495	0.58	**0.02**	0.001	0.53	0.10	0.001	0.59	**0.02**	0.001	0.55	0.10	0.001
Diploma-BA	0.80	**0.21**	0.389	0.68	**0.03**	0.001	0.63	0.12	0.017	0.70	**0.03**	0.001	0.67	0.13	0.037
Over BA	0.27	**0.12**	0.001	0.20	**0.03**	0.001	0.27	0.10	0.001	0.31	**0.04**	0.001	0.29	0.11	0.001
AIC		1.23			17.56			5.57			13.39			5.55	
Log likelihood		-420.46			-6135.85			-1937.67			-4668.97			-1925.34	
SSE		144.930			23.526			0.260			1.099			0.220	

OR: odds ratio; **SE**: standard error; **IRR**: incidence rate ratio; **AIC**: Akaiki information criteria; **SSE**: sum of squared error

As expected results of logistic regression was not satisfying. This model failed to detect significance of education. In addition, effects of province, and middle age group were of marginal significance. SSE corresponding to this model was considerably higher than the other models.

Variables affected respondents to reveal the number of known alcoholics were province, gender, age groups

and education levels (Table 2). However, marital status was not informative. These four variables were retained significant in all four models. Based on ZINB model, Kermaninans' were about 25% less likely to know alcoholics than those from Fars (OR=0.74, *P*=0.015). Furthermore, women were 70% less likely to know alcoholics than male. Respondents in less than 30 years old and those with low education were more likely to know or reveal alcoholics. Increasing in age and education level was associated with reduction of revealing information about the number of alcoholics, people knew.

Factors influence knowing methadone users

With nearly 50% zero response, the NB explained the observed data better than P (*P*<0.001). However, the ZINB model was not preferred over the NB (*P*=0.482). Additionally there was no difference in the AIC (3.93 vs. 3.92) and SSE (0.28 vs. 0.27) of ZINB and NB. Figure 5 shows that the performance of both NB and ZINB models, in terms of prediction of zeros, were almost the same. Therefore the simpler model (NB) was chosen as the best model.

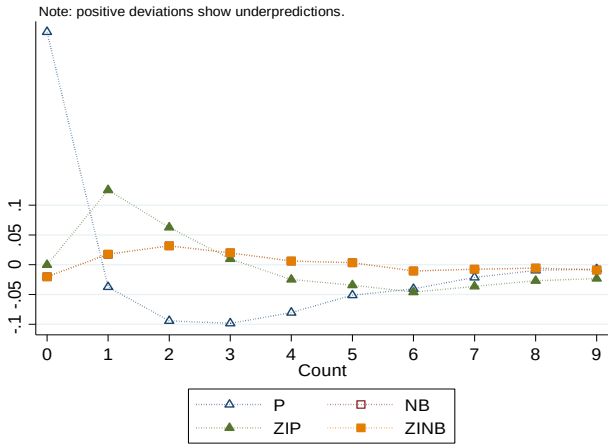

Figure 5: Comparisons among observed versus predicted probabilities among count models for methadone

Table 3: Influence of demographic characteristics of respondents about number of methadone users in their network considering zero inflated negative binomial as gold standard; the bold figures mean SEs estimated with bias (>15% difference with the gold model)

	Logistic			Poisson			Negative binomial			Zero inflated Poisson			Zero inflated Negative binomial		
Variables	**0R**	**SE**	***P* value**	**IRR**	**SE**	***P* value**	**IRR**	**SE**	***P* value**	**IRR**	**SE**	***P* value**	**IRR**	**SE**	***P* value**
Province															
Fars		1.00			1.00			1.00			1.00			1.00	
Kerman	1.59	**0.25**	0.003	2.18	**0.10**	0.001	2.06	0.32	0.001	1.61	**0.08**	0.001	2.03	**0.38**	0.001
Gender															
Male		1.00			1.00			1.00			1.00			1.00	
Female	1.15	**0.19**	0.374	0.72	**0.03**	0.001	0.9	0.15	0.537	0.69	**0.03**	0.001	0.90	0.16	0.529
Marital status															
Single		1.00			1.00			1.00			1.00			1.00	
Ever married	1.38	**0.27**	0.097	0.88	**0.05**	0.018	1.01	0.19	0.964	0.80	**0.04**	0.001	1.01	0.19	0.971
Age (yr)															
15-29		1.00			1.00			1.00			1.00			1.00	
30-39	0.89	0.18	0.581	0.99	**0.06**	0.921	0.88	0.18	0.550	0.92	**0.06**	0.148	0.88	0.18	0.555
≥40	0.82	0.20	0.413	1.06	0.07	0.358	0.96	0.22	0.854	1.04	**0.07**	0.560	0.96	0.22	0.862
Education															
Under diploma		1.00			1.00			1.00			1.00			1.00	
Diploma	1.09	**0.26**	0.717	0.61	**0.04**	0.001	0.59	0.14	0.022	0.57	**0.03**	0.001	0.59	0.14	0.022
Diploma-BA	1.17	**0.28**	0.521	0.49	**0.03**	0.001	0.56	0.13	0.012	0.46	**0.03**	0.001	0.56	0.13	0.012
Over BA	0.48	**0.22**	0.106	0.24	**0.04**	0.001	0.22	0.10	0.001	0.33	**0.07**	0.001	0.22	0.10	0.001
AIC			1.38			10.28			3.92			7.78			3.93
Log likelihood			-475.02			-3590.37			-1358.93			-2708.12			-1358.92
SSE			170.20			54.11			0.27			2.01			0.28

OR: odds ratio; SE: standard error; **IRR**: incidence rate ratio; **AIC**: Akaiki information criteria; **SSE**: sum of squared error

Based on NB model, Kermaninans, relative to respondents in Fars, was 2.06 times more likely to know or reveal information about the number of methadone users they knew. Education level was associated with the outcome as well, where less educated people knew more methadone users than those in the other categories. In particular people with postgraduate degree, relative to those with the lowest degree, were about 80% less likely to reveal or know methadone users. However, effects of gender and marital status were controversial. Gender and marital status did not retain in NB but did in P and ZIP. As was shown in Table 3, SE's corresponding to P and ZIP models are around four times smaller than NB and ZINB models. This explains these findings.

Again, results of logistic model were not satisfying. This model produced very large SSE. Regarding the variables retained in the model, significance of education levels was lost.

Factors influence knowing FSW

Regarding the FSW, the LRT of over-dispersion comparing the NB to the P yielded a *P*<0.001. However, ZINB model did not preferred over the NB (*P*=0.175). AIC and SSE value showed poor fit of logistic and P

models. Figure 6 shows that both, the ZINB and NB fit the data well.

The SEs for the P and ZIP were smaller than that of NB (Table 4). However, results of NB and ZINB were nearly equal because they provided the same fit to the data according the formal Vuong test.

Based on the NB model (gold model), people in Kerman were about 42% less likely to know FSWs than those in Fars with a *P*-value of 0.002.

The effect of province was of marginal significance in ZIP and ZINB models. Gender was significant in all models, where females were about 80% less likely to know FSW than male (in NB model). Marital status was not informative in any of the fitted models. In addition, all models confirmed that those in middle age and old age group, and well educated people were less likely to know FSWs.

Analyzing FSW data with logistic model, in terms of effective variables in final model, results were exactly the same as NB.

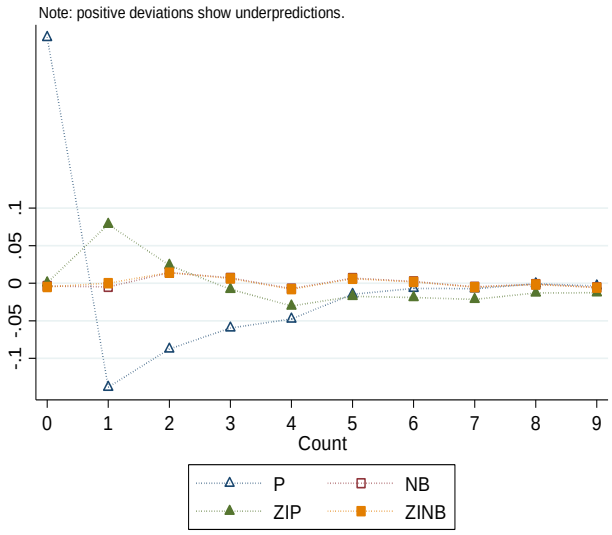

Figure 6: Comparisons among observed versus predicted probabilities among count models for female sex worker

Table 4: Influence of demographic characteristics of respondents about number of female sex workers in their network considering zero inflated negative binomial as gold standard; the bold figures mean SEs estimated with bias (>15% difference with the gold model)

	Logistic			Poisson			Negative binomial			Zero inflated Poisson			Zero inflated Negative binomial		
Variables	**0R**	**SE**	***P* value**	**IRR**	**SE**	***P* value**	**IRR**	**SE**	***P* value**	**IRR**	**SE**	***P* value**	**IRR**	**SE**	***P* value**
Province															
Fars		1.00			1.00			1.00			1.00			1.00	
Kerman	0.62	**0.11**	0.006	0.70	**0.04**	0.001	0.58	0.10	0.002	0.89	**0.06**	0.061	0.69	**0.14**	0.066
Gender															
Male		1.00			1.00			1.00			1.00			1.00	
Female	0.43	**0.07**	0.001	0.28	**0.02**	0.001	0.22	0.04	0.001	0.44	**0.03**	0.001	0.25	**0.05**	0.001
Marital status															
Single		1.00			1.00			1.00			1.00			1.00	
Ever married	0.79	0.16	0.253	0.93	**0.07**	0.326	0.82	0.18	0.375	1.01	**0.07**	0.863	0.96	**0.23**	0.855
Age (yr)															
15-29		1.00			1.00			1.00			1.00			1.00	
30-39	0.71	0.16	0.131	0.76	**0.06**	0.001	0.64	0.15	0.065	0.81	**0.07**	0.011	0.65	0.16	0.083
≥40	0.43	**0.12**	0.003	0.40	**0.04**	0.001	0.34	0.10	0.001	0.60	**0.07**	0.001	0.34	0.10	0.001
Education															
Under diploma		1.00			1.00			1.00			1.00			1.00	
Diploma	0.46	0.12	0.002	0.35	**0.03**	0.001	0.24	0.06	0.001	0.48	**0.04**	0.001	0.24	0.07	0.001
Diploma-BA	0.55	**0.14**	0.019	0.42	**0.03**	0.001	0.31	0.08	0.001	0.53	**0.04**	0.001	0.29	0.09	0.001
Over BA	0.23	**0.13**	0.008	0.13	**0.04**	0.001	0.14	0.08	0.001	0.27	0.09	0.001	0.13	0.07	0.001
AIC			1.21			6.08			2.83			4.42			2.84
Log likelihood			-415.87			-2120.70			-977.52			-1529.69			-975.36
SSE			143.22			50.73			0.02			0.55			0.02

OR: odds ratio; **SE**: standard error; **IRR**: incidence rate ratio; **AIC**: Akaiki information criteria; **SSE**: sum of squared error

Discussion

The efficacy of the P, NB, ZIP, ZINB and logistic regression was investigated for modeling of count skewed data. Three outcome variables with different percentages of zero were modeled.

In our analyses, we have seen that NB and ZINB models gave the best fit to the outcomes. P and ZIP did not provide adequate fit to the outcomes. This was not against our expectation, as we have seen great over-dispersion in all outcomes. Additionally the results of logistic regression were not satisfying in outcomes, with low and moderate zero values, due to very large SSEs. For the last outcome (FSWs) since third quarter of the data was one, at most 25% of the data analyzed in logistic and NB models were different. This partially explains why the results were so closed. However, even in this scenario SSE corresponding to logistic model was considerably higher than NB model.

In accordance to the usual practice, we applied LRT and Vuong tests to compare nested and non-nested models.

It has been suggested that "NB is capable of predicting large percent of zero when the count range is not too large, even better than ZINB" [8]. This was consistent with our findings. In our data, the distribution of the number of methadone users and FSWs was not that large. Medians of these two variables were one and zero, and third quartiles were three and one respectively. Distribution of number of alcoholics was more diverge with median and third quartiles at three and eight.

In a study by Sileshi the minimum and the maximum of zero percentage in 11 data sets were 50 and 90 respectively. Altogether the NB had the best fit in this study [9]. This indicates that rarity cannot be the only criterion for application of zero inflated models. Wenger et al. expressed that unlike the past that zero-inflated models had been advised for rare events, studies proved that NB can also fit the apparently zero-inflated data well[10].

Performance of regression models was compared using 12 count data sets. The zero percentage in these 12 data sets ranged from 17% to 91%. In ten data sets NB provided the best fit. The minimum and maximum of zero percentage in these ten data sets were 17% and 91% respectively. The results confirmed the presence of over-dispersion and excess zero in the data sets[11].

Analyzing environmental data, two count data sets were analyzed. In the first data the percentage of zeroes was roughly 41%. The AICs of the fitted models were close, but the best model was still the P. Authors noted that in their data set the P was able to predict proportion of zeros and ZIP was not needed. In the second data set, of 1386 observations, 54% had zero count. The ZINB provided the best fit as this model was able to take into account both excess of zeroes and over-dispersion [12].

On the other hand, modeling young driver motor vehicle crashes, the outcome variable involved 90% zero counts. Authors have seen that the NB and ZIP models provided adequate fits to the motor vehicle crash data. It has been commented that, when over-dispersion exists, NB and ZIP models are potential alternatives for P [13]. However, comparison of these models with ZINB was not performed at all.

There are examples with moderate percentage of zero, in which performance of P and ZIP models were satisfying. In a study by Cheung et al., with around 13% zero counts in the outcome, the ZIP model provided the best fit. In this study the ZINB model did not show any over-dispersion after the extra zeros had been taken into account [14].

There are many studies in which performance of all the four count models are not study simultaneously. For example a literature review by Preisser and his colleagues sought to review all published research articles in the dental literature that used ZIP or ZINB models to analyze caries experience. From 15 articles, six studies performed both ZIP and ZINB and to compare them in two of the studies tests (LRT, AIC, BIC) and graphical approaches and in the others one of this approaches was used.

Other aspect of our study was to address characteristics of respondents who were more likely to reveal sensitive information. We have seen that young respondents, male, and those with low education were more likely to know (or reveal information) people from hidden sub groups. In addition, Kermaninans were less likely to know (or reveal information about) alcoholics and FSWs than those in Fars. The opposite was true in the case of methadone users. In a similar study, Shelley et al. (1995) found that HIV+ patients were not willing to share information about their HIV status to all members of their network. Medical personnel and support group members had the most knowledge of HIV status. Gender and ethnicity also influenced this knowledge [15].

Several issues might explain our findings. The first explanation might be cultural differences. People in Kerman might feel free to reveal information about drug using, but in Fars to reveal sexual and alcohol related behaviors. This mean that even in two close provinces, degree of conservatism about different issues might not be the same. It has been suggested that the pattern of alcohol use varies in different societies, as cultural values affect the seriousness of such behaviors. The second reason might be that the number of FSWs and alcoholics in Fars was greater than Kerman, while the number of those receives methadone in Kerman, was larger than Fars province. Our size estimation results confirmed this hypothesis (confidential data submitted to the IMHME).

Our study has several limitations. Firstly, we only consider three outcomes. Modeling of outcomes with less or higher percentage of zeros may have different goodness of fit. In addition, our literature review showed that over-dispersion, excess zero, and range of data are other issues affected performance of count regression models. Therefore, future studies should be designed to address impact of different combination of these issues on performance of models. There is also a need to use the results of hurdle models and finite mixture models as an alternative to zero-inflated models in future investigations.

One last limitation of this study was that all data were self-reported. We simply asked respondents about number of people they know in each hidden group. No external data was available to check the accuracy of responses. However, our purpose in this manuscript was not to provide accurate size of hidden groups, but to address art of statistical modeling to explore the impact of demographic characteristics of respondents on their attitude to reveal sensitive issues.

Conclusion

We noticed that P might not be adequate for modeling of count data. In the cases that over-dispersion or excess zero exists; alternative count regression models should be explored. In such scenarios, application of P leads to small SEs. Consequently variables with no impact on the outcome would become significant. Furthermore, we have seen that demographic characteristics of respondents affected their knowledge and/ or tendency to reveal sensitive behaviors of those in their network.

Acknowledgments

The authors would like to thank Iranian Ministry of Health due to the data. This article is the result of a biostatistics master of science's thesis which its research project was approved at Kerman Medical Student Research Committee.

Conflict of interest statement

The authors declare that they have no conflicts of interest.

Funding

The project has been funded by Mental and Social Health, Addiction Department of Ministry of Health and proposed by the ministry.

References

1. Shokoohi M, Baneshi MR, Haghdoost A-a. Size estimation of groups at high risk of HIV/AIDS using network scale up in Kerman, Iran. *Int J Prev Med*.3(7):471.
2. Zaninotto P, Falaschetti E. Comparison of methods for modeling a count outcome with excess zeros: application to Activities of Daily Living (ADL-s). *J Epidemiol Community Health*.65(3):205.
3. Famoye F, Singh KP. Zero-inflated generalized Poisson regression model with an application to domestic violence data. *J Data Sci*. 2006;4(1):117-130.
4. Karazsia BT, van Dulmen MHM. Regression models for count data: Illustrations using longitudinal predictors of childhood injury. *J Pediatr Psychol*. 2008;33(10):1076.
5. Slymen DJ, Ayala GX, Arredondo EM, Elder JP. A demonstration of modeling count data with an application to physical activity. *Epidemiol Perspect Innov*. 2006;3(1):3.
6. Yesilova A, Kaya Y, Kaki B, Kasap I. Analysis of plant protection studies with excess zeros using zero-inflated and negative binomial hurdle models. *GU J Sci*.23(2):131-136.
7. Yesilova A, Kaydan MB, Kaya Y. Modeling insect-egg data with excess zeros using zero-inflated regression models. *Hacettepe Journal of Mathematics and Statistics*. 2010;39(2):273-282.
8. Xia Y, Morrison-Beedy D, Ma J, Feng C, Cross W, Tu X. Modeling Count Outcomes from HIV Risk Reduction Interventions: A comparison of competing statistical models for count responses. *AIDS Res Treat*.2012. 2012:593569.
9. Sileshi G. Selecting the right statistical model for analysis of insect count data by using information theoretic measures. Bu*ll Entom*ol Res. 2006;96(05):479-488.
10. Wenger SJ, Freeman MC. Estimating species occurrence, abundance, and detection probability using zero-inflated distributions. *Eco Soc America*. 2008;89(10):2953-2959.
11. Sileshi G. The excess-zero problem in soil animal count data and choice of appropriate models for statistical inference. *Pedobiologia*. 2008;52(1):1-17.
12. Viviano LCM, Muggeo VMR, Lovison G. Using zero-inflated models to analyze environmental data sets with many zeroes. 2005. Available from: http://cab.unime.it/mus/385/1/Using_Zero-inflated_Models_to_Analyze_Environmental_Data_Sets_with_Many_Zeroes.pdf.
13. Lee AH, Stevenson MR, Wang K, Yau KKW. Modeling young driver motor vehicle crashes: data with extra zeros. *Accid Anal Prev*. 2002;34(4):515-521.
14. Cheung YB. Zero • inflated models for regression analysis of count data: a study of growth and development. *Stat Med*. 2002;21(10):1461-1469.
15. Johnsen EC, Bernard HR, Killworth PD, Shelley GA, McCarty C. A social network approach to corroborating the number of AIDS/HIV+ victims in the US. *Soc Networks*. 1995;17(3-4):167-187.