Communications for Statistical Applications and Methods
2018, Vol. 25, No. 2, 173–184

https://doi.org/10.29220/CSAM.2018.25.2.173
Print ISSN 2287-7843 / Online ISSN 2383-4757

Weighted zero-inflated Poisson mixed model with an application to Medicaid utilization data

Sang Mee Lee[1,a], Theodore Karrison[a], Robert S. Nocon[b], Elbert Huang[b]

[a]Department of Public Health Sciences, University of Chicago, USA;
[b]Department of Medicine, University of Chicago, USA

Abstract

In medical or public health research, it is common to encounter clustered or longitudinal count data that exhibit excess zeros. For example, health care utilization data often have a multi-modal distribution with excess zeroes as well as a multilevel structure where patients are nested within physicians and hospitals. To analyze this type of data, zero-inflated count models with mixed effects have been developed where a count response variable is assumed to be distributed as a mixture of a Poisson or negative binomial and a distribution with a point mass of zeros that include random effects. However, no study has considered a situation where data are also censored due to the finite nature of the observation period or follow-up. In this paper, we present a weighted version of zero-inflated Poisson model with random effects accounting for variable individual follow-up times. We suggested two different types of weight function. The performance of the proposed model is evaluated and compared to a standard zero-inflated mixed model through simulation studies. This approach is then applied to Medicaid data analysis.

Keywords: emergency department, Health care utilization, weight function, zero-inflated model

1. Introduction

Health care utilization data are often analyzed to address critical questions about health care service and delivery, such as resource utilization and planning, allocation of services and the evaluation of patient outcomes. Such analysis is of increasing importance for policy makers and health care institutions to improve the quality of patient care. Measurement of utilization includes the frequency of visits to medical providers, visits to emergency department (ED), days spent in a hospital, and use of prescription medication. Such count data are frequently found overdispersed with heavy tails and are often multi-modal with excess zeroes. For example, few patients use a service multiple times, whereas the majority report no utilization in a specific time period. Traditional approaches for overdispersed count data with extra variation have been the use of zero-inflated models, that include zero-inflated Poisson (ZIP) (Lambert, 1992) and zero-inflated negative binomial (ZINB) (Ridout *et al.*, 1998). These models were developed by assuming that the outcome variable contains a mixture of a point mass at zero and a count distribution. For a comprehensive review of zero-inflated models, see Ridout *et al.* (1998). However, a hurdle model (Mullahy, 1986) has been independently developed for count data with excessive zeros assuming all zeros are from one structural source rather than two sources

[1] Corresponding author: Department of Public Health Sciences, University of Chicago, 5841 S. Maryland Ave. MC 2000, IL 60637, USA. E-mail: slee@health.bsd.uchicago.edu

Published 31 March 2018 / journal homepage: http://csam.or.kr

(both structural and sampling zeros) as assumed in the ZIP models. Thus, positive count data (non-zero) are assumed as either truncated Poisson or truncated negative binomial distribution whereas logistic regression part predicts all zeros in a hurdle model. The model choice between zero-inflated and hurdle models of the distinction between structural and sampling zeros, may be subtle in practice (Monod, 2014). In our application, we develop an approach based on a case in which structural zeros cannot be distinguished from random zeros; therefore, we choose to focus on zero-inflated models.

Data on health care use often have a multilevel structure where patients are typically nested within physicians, hospitals, and geographic regions. Consequently, intra-cluster correlation and heterogeneity between clusters must be considered. In such cases, random effects are commonly incorporated into the zero-inflated model. Hall (2000) proposed ZIP with a random intercept to account for the within-subject dependence in the Poisson state. Yao and Lee (2001) introduced a pair of uncorrelated random effects to both zero and Poisson components. Min and Agresti (2005) suggested linking the two components by the joint distribution of the random effects. Lee *et al.* (2006) and Moghimbeigi *et al.* (2008) extended to a three-level hierarchical model using normally-distributed random effects.

In addition to a clustered structure, health care utilization data are often censored due to the finite nature of the observation period or follow-up. For instance, when utilization data are collected in long-term observational or randomized clinical trials, complete data are not available for some patients because they are not followed until the endpoint of interest or they enter in the middle of the study. Thus, patients with shorter follow-up should be treated differently from patients who have data for the entire study period because the incidence rate of utilization would otherwise be underestimated. Several authors have proposed modified zero-inflated models to handle variable follow-up (Emerson *et al.*, 1993; Hsu, 2005, 2007). They adapted a weight function for the simple fact that the likelihood of observing a greater number of utilizations is higher for patients who are followed for longer periods of time. However, no study has considered this in conjunction with multi-level data. This provided the motivation to modify zero-inflated models with random effects by incorporating a variable follow-up period.

In this paper, we present a weighted ZIP model with random effects for clustered count data with excess zeros. The proposed model accounts for variable individual follow-up times using a weight function that can be estimated by several approaches. We focus on Poisson models, but the methods can easily be extended to other count distributions, such as the negative binomial.

The paper is organized as follows. In Section 2 we define the weighted ZIP mixed model. In Section 3, we describe the estimation procedures. In Section 4 simulation studies are conducted to show model performance. Section 5 applies the models to an analysis of Medicaid data. Discussion and future research are provided in Section 6.

2. Weighted zero-inflated Poisson mixed model

We consider incorporating the duration of follow-up in ZIP mixed model through a weight function. Let Y_{ij} be the outcome and t_{ij} be the follow-up time for the j^{th} subject within the i^{th} cluster ($i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n_i$). The probability of observing events is denoted as $\pi_{ij} = \Pr(Y_{ij} > 0)$ and the total number of individuals $n = \sum_{i=1}^{m} n_i$. The weight function, $w(t_{ij})$, where $0 < w(t_{ij}) \leq 1$, is assumed to be an increasing function. $w(t_{ij}) = 1$ indicates the data is completely observed over the study period whereas $w(t_{ij}) < 1$ indicates the information is partially observed. This accounts for the increased probability of observing an event for individuals with longer follow-ups than those with shorter periods of time. We assume that the probability of observing events over time t_{ij}, $\Pr(Y_{ij} > 0; t_{ij})$, is equal to $w(t_{ij}) \Pr(Y_{ij} > 0)$.

The weighted ZIP mixed model is

$$\Pr(Y = y_{ij}; t_{ij}) = \begin{cases} 1 - w(t_{ij})\pi_{ij} + w(t_{ij})\pi_{ij}e^{-\lambda_{ij}}, & y_{ij} = 0, \\ w(t_{ij})\pi_{ij}\dfrac{\lambda_{ij}^{y_{ij}}e^{-\lambda_{ij}}}{y_{ij}!}, & y_{ij} > 0 \end{cases} \tag{2.1}$$

and

$$\text{logit}(\pi_{ij}) = \xi_{ij} = z'_{ij}\alpha + u_i, \qquad \log\left(\lambda_{ij}\right) = \eta_{ij} = x'_{ij}\beta + v_i,$$

where z_{ij} and x_{ij} are respectively vectors of covariates for the logistic and the Poisson components, and α and β are the corresponding vectors of regression coefficients. The u_i and v_i denote the random effects and are assumed to be independent and normally distributed with mean 0 and variance σ_u^2 and σ_v^2, respectively. In this paper, we consider two weights functions. One is a uniform weight function (denoted by W-ZIP^u), i.e., $w(t) = t/T$ where T is the complete observation time. The other is an exponential weight function (denoted by W-ZIP^e), i.e., $w(t) = (1 - e^{-\lambda t})/(1 - e^{-\lambda T})$, where λ is a constant hazard and is estimated from the data using an exponential survival function. Their performance is explored in the simulation studies.

3. Model estimation

To ensure convergence in estimation of parameters and random effects the penalized log-likelihood is given by $l = l_1 + l_2$, where

$$\begin{aligned} l_1 = &\sum_{\{ij:y_{ij}=0\}} \log\left[1 + \exp(\xi_{ij}) - w(t_{ij})\exp(\xi_{ij}) + w(t_{ij})\exp(\xi_{ij})\exp(-\exp(\eta_{ij}))\right] - \log\left[1 + \exp(\xi_{ij})\right] \\ &+ \sum_{\{ij:y_{ij}>0\}} \log(w(t_{ij})) + \xi_{ij} - \log\left[1 + \exp(\xi_{ij})\right] + y_{ij}\eta_{ij} - \exp(\eta_{ij}) - \log(y_{ij}!), \\ l_2 = &-\frac{1}{2}\left(m\log\left(2\pi\sigma_u^2\right) + \sigma_u^{-2}u'u + m\log\left(2\pi\sigma_v^2\right) + \sigma_v^{-2}v'v\right) \end{aligned}$$

with l_1 being the log-likelihood when the random effects are conditionally fixed, and l_2 being the penalty. The complete data log-likelihood l_c is constructed as $l_c = l_\xi + l_\eta$ with

$$\begin{aligned} l_\xi &= \sum_{i,j}\psi_{ij}\log\left(1+\exp(\xi_{ij})-w(t_{ij})\exp(\xi_{ij})\right)+\left(1-\psi_{ij}\right)\xi_{ij}-\log\left(1+\exp(\xi_{ij})\right)-\frac{1}{2}\left(m\log\left(2\pi\sigma_u^2\right)+\sigma_u^{-2}u'u\right), \\ l_\eta &= \sum_{i,j}(1-\psi_{ij})\left(y_{ij}\eta_{ij} - \exp(\eta_{ij})\right) - \frac{1}{2}\left(m\log\left(2\pi\sigma_v^2\right) + \sigma_v^{-2}v'v\right), \end{aligned}$$

where ψ_{ij} is a latent variable indicating whether y_{ij} comes from zero ($\psi_{ij} = 1$) or non-zero($\psi_{ij} = 0$) state. The complete log-likelihood l_c can be easily maximized by maximizing the l_ξ and l_η separately. With EM algorithm, ψ_{ij} is estimated by its conditional expectation $\psi_{ij}^{(k)}$ under the current estimates $\alpha^{(k)}, \beta^{(k)}, u^{(k)}$, and $v^{(k)}$. Here

$$\psi_{ij}^{(k)} = \begin{cases} \left[1 + \dfrac{w(t_{ij})\exp\left(-\exp\left(x'_{ij}\beta^{(k)} + v_i^{(k)}\right)\right)}{1 - w(t_{ij}) + \exp\left(-t'_{ij}\alpha^{(k)} - u_i^{(k)}\right)}\right]^{-1}, & y_{ij} = 0, \\ 0, & y_{ij} > 0. \end{cases}$$

Details are given in the Appendix A.

Table 1: Bias, MSE, and Cov for ZIP and W-ZIP mixed models (uniform weight function)

		ZIP mixed			W-ZIP^u mixed			W-ZIP^e mixed		
		Bias	MSE	Cov†	Bias	MSE	Cov	Bias	MSE	Cov
	α_0	−1.292	1.735	0.000	−0.037	0.268	0.930	−0.194	0.297	0.912
	α_1	−0.673	0.627	0.662	0.047	0.966	0.940	0.143	0.867	0.956
$n = 250$	σ_u^2	−0.046	0.007	0.282	0.098	0.080	0.750	0.067	0.058	0.790
$m = 10$	β_0	0.008	0.015	0.952	0.010	0.015	0.952	0.022	0.017	0.924
	β_1	0.002	0.017	0.938	0.000	0.017	0.944	−0.011	0.017	0.940
	σ_v^2	−0.010	0.005	0.292	−0.010	0.005	0.300	−0.008	0.004	0.282
	α_0	−1.282	1.714	0.000	−0.001	0.287	0.918	−0.127	0.337	0.900
	α_1	−0.654	0.615	0.670	0.093	0.982	0.948	0.019	1.057	0.926
$n = 250$	σ_u^2	−0.058	0.007	0.446	0.080	0.077	0.854	0.050	0.062	0.886
$m = 25$	β_0	−0.002	0.026	0.942	−0.001	0.026	0.940	0.004	0.029	0.930
	β_1	0.009	0.015	0.944	0.007	0.014	0.948	−0.001	0.017	0.948
	σ_v^2	−0.002	0.010	0.490	−0.003	0.010	0.496	−0.004	0.010	0.456
	α_0	−1.289	1.694	0.000	0.000	0.138	0.928	−0.145	0.149	0.892
	α_1	−0.646	0.520	0.482	−0.017	0.509	0.914	−0.005	0.413	0.950
$n = 500$	σ_u^2	−0.066	0.006	0.154	0.030	0.023	0.730	0.016	0.021	0.692
$m = 10$	β_0	0.005	0.010	0.964	0.006	0.010	0.962	0.005	0.013	0.942
	β_1	0.001	0.007	0.956	−0.000	0.007	0.956	0.003	0.008	0.950
	σ_v^2	−0.005	0.004	0.306	0.005	−0.004	0.306	−0.004	0.004	0.298
	α_0	−1.299	1.728	0.000	−0.026	0.145	0.924	−0.155	0.139	0.900
	α_1	−0.631	0.508	0.470	0.029	0.484	0.934	0.008	0.401	0.938
$n = 500$	σ_u^2	−0.078	0.007	0.124	0.009	0.022	0.874	0.000	0.016	0.846
$m = 25$	β_0	0.008	0.024	0.926	0.009	0.024	0.928	−0.010	0.025	0.914
	β_1	−0.006	0.007	0.950	−0.007	0.007	0.954	−0.003	0.008	0.936
	σ_v^2	−0.004	0.008	0.504	−0.004	0.008	0.502	0.009	0.011	0.474

MSE = mean square error; Cov = coverage probability; ZIP = zero-inflated Poisson; W-ZIP = weighted ZIP.
†: Coverage probability is the proportion of times the estimated 95% confidence interval contains the true value.

4. Simulation

Simulation studies were conducted to evaluate the performance of the proposed weighted ZIP mixed models. Each of 500 data sets was generated from W-ZIP mixed model in (2.1). We consider one common covariate for both logistic and Poisson components $z_{ij} = x_{ij}$ generated from Uniform $(0, 1)$. The true values of the regression coefficients are assumed to be $\alpha' = (1, 1)$ and $\beta' = (1, 2)$. The variances of random effects for the logistic u_i and Poisson part v_i are set as $\sigma_u^2 = 0.1$, $\sigma_v^2 = 0.2$. To mimic the Medicaid data analysis, we simulated observation time periods t_{ij} from an exponential distribution with a parameter 0.1 and set the maximum follow-up at 12 month. As a result, the mean observation follow-up time was 6.8. The response y_{ij} is obtained from weighted ZIP mixed model with either uniform or exponential weight. Two sample sizes $n = 250, 500$ with $m = 10, 25$ clusters are considered in this paper.

Tables 1 and 2 presents the results of evaluating the proposed W-ZIP mixed models compared to ZIP mixed model. The bias, mean square error (MSE) and coverage probability are used to compare the performance of the models. The W-ZIP mixed models performs well in estimating all the regression coefficients whereas estimates of logistic part in ZIP mixed models were biased with larger MSE and smaller coverage rate compared with W-ZIP models. In particular, the estimated 95% confidence interval for $\hat{\alpha}_0$ in ZIP model never contained the true coefficient. The coverage probability for σ_u^2 and σ_v^2 across all models was not close to the nominal confidence level but it increases as the cluster size increases from 10 to 25. As expected, the bias and MSE decrease with increasing sample size.

In addition, the sensitivity of the proposed model was evaluated given data with outliers. We

Table 2: Bias, MSE, and Cov for ZIP and W-ZIP mixed models (exponential weight function)

		ZIP mixed			W-ZIP^u mixed			W-ZIP^e mixed		
		Bias	MSE	Cov†	Bias	MSE	Cov	Bias	MSE	Cov
	α_0	−1.224	1.573	0.004	0.094	0.338	0.944	−0.064	0.244	0.942
	α_1	−0.620	0.590	0.710	0.034	1.058	0.956	0.078	0.911	0.958
$n = 250$	σ^2_u	−0.042	0.009	0.226	0.116	0.099	0.764	0.081	0.070	0.764
$m = 10$	β_0	0.010	0.017	0.944	0.012	0.017	0.940	0.013	0.015	0.942
	β_1	0.004	0.017	0.938	0.002	0.017	0.946	−0.004	0.016	0.928
	σ^2_v	−0.005	0.004	0.348	−0.005	0.004	0.336	−0.010	0.004	0.284
	α_0	−1.207	1.529	0.000	0.132	0.616	0.928	0.012	0.314	0.948
	α_1	−0.621	0.602	0.686	0.081	1.628	0.942	0.025	0.905	0.954
$n = 250$	σ^2_u	−0.063	0.007	0.392	0.059	0.065	0.878	0.078	0.075	0.868
$m = 25$	β_0	−0.014	0.027	0.954	−0.013	0.027	0.954	0.002	0.031	0.910
	β_1	0.008	0.015	0.942	0.007	0.015	0.944	−0.005	0.014	0.946
	σ^2_v	−0.003	0.010	0.500	−0.003	0.010	0.498	−0.001	0.010	0.484
	α_0	−1.197	1.466	0.000	0.075	0.155	0.938	−0.018	0.126	0.946
	α_1	−0.650	0.520	0.440	0.071	0.500	0.946	0.023	0.421	0.958
$n = 500$	σ^2_u	−0.065	0.006	0.144	0.040	0.034	0.732	0.027	0.026	0.718
$m = 10$	β_0	0.003	0.013	0.930	0.005	0.013	0.930	0.010	0.012	0.944
	β_1	0.003	0.007	0.944	0.001	0.007	0.944	−0.006	0.007	0.954
	σ^2_v	−0.001	0.004	0.312	−0.001	0.004	0.310	0.002	0.004	0.308
	α_0	−1.197	1.463	0.000	0.087	0.123	0.962	0.016	0.126	0.952
	α_1	−0.634	0.485	0.452	0.083	0.407	0.960	0.005	0.403	0.946
$n = 500$	σ^2_u	−0.075	0.007	0.156	0.016	0.028	0.850	0.030	0.029	0.800
$m = 25$	β_0	−0.004	0.025	0.924	−0.003	0.025	0.926	0.002	0.022	0.940
	β_1	0.006	0.007	0.942	0.005	0.007	0.946	−0.003	0.006	0.954
	σ^2_v	−0.006	0.009	0.486	−0.006	0.009	0.482	−0.004	0.009	0.488

MSE = mean square error; Cov = coverage probability; ZIP = zero-inflated Poisson; W-ZIP = weighted ZIP.
†: Coverage probability is the proportion of times the estimated 95% confidence interval contains the true value.

Table 3: Evaluation for W-ZIP mixed models given data with outliers

			W-ZIP^u mixed			W-ZIP^e mixed		
			Bias	MSE	Cov†	Bias	MSE	Cov
		α_0	0.060	0.266	0.950	−0.054	0.237	0.952
	10% outliers	α_1	0.088	1.058	0.956	0.059	0.941	0.958
		β_0	0.014	0.047	0.912	0.013	0.047	0.912
$n = 250$		β_1	−0.012	0.022	0.914	−0.011	0.022	0.914
$m = 10$		α_0	0.088	0.336	0.950	−0.025	0.291	0.948
	20% outliers	α_1	0.045	0.896	0.962	0.018	0.798	0.964
		β_0	0.000	0.070	0.904	−0.001	0.070	0.904
		β_1	−0.023	0.030	0.918	−0.023	0.030	0.918

MSE = mean square error; Cov = coverage probability; ZIP = zero-inflated Poisson; W-ZIP = weighted ZIP.
†: Coverage probability is the proportion of times the estimated 95% confidence interval contains the true value.

used the same simulation settings described earlier and randomly selected 10% or 20% of outcomes. Then outliers were generated from weight ZIP model in (2.1) with $u_i \sim N(0, 1)$ and $v_i \sim N(0, 2)$ for outliers. Table 3 shows that all W-ZIP models have low bias and MSE for estimating β and the coverage probabilities for all estimates were close to the nominal 95%.

5. Application to Medicaid data

To demonstrate the approach with real study data, we applied it to a study of federally-funded community health centers (HCs) funded by the Bureau of Primary Health Care (BPHC). This study was

Table 4: Means of patients' characteristics before and after the propensity score matching

		Unmatched (n = 103,379)		Matched (n = 40,122)		% Balance
		HC	Non-HC	HC	Non-HC	Improvement
		n = 20,061	n = 83,318	n = 20,061	n = 20,061	Mean Diff.
Age		31.4	33.0	31.4	31.5	94.9
Female		0.816	0.731	0.816	0.813	96.3
Race	White	0.186	0.509	0.186	0.189	99.1
	Black	0.539	0.283	0.539	0.538	99.6
	Hispanic	0.134	0.073	0.134	0.135	99.2
	Others	0.141	0.135	0.141	0.138	42.9
Medicaid eligibility group[a]	Cash assistance	0.166	0.129	0.166	0.165	96.8
	Blind/disabled	0.206	0.274	0.206	0.208	96.9
	Medical need	0.409	0.439	0.409	0.386	23.7
	Poverty	0.279	0.221	0.279	0.291	79.1
	Aged	0.006	0.011	0.006	0.006	94.4
	Others	0.365	0.403	0.365	0.371	82.4
TANF eligible[b]		0.057	0.027	0.057	0.056	99.0
Restricted benefits[c]		0.107	0.040	0.107	0.111	92.9
Delivery during the year		0.079	0.044	0.079	0.076	91.3
CDPS risk score[d]		0.805	1.028	0.805	0.811	97.1

[a] : Enrollees may be have more than one eligibility category assigned over the course of the year. Eligibility categories are grouped from original Medicaid Analytic eXtract data.
[b] : Enrollee is eligible for Temporary Aid For Needy Families (TANF) program in any month during the data year.
[c] : Enrollee is eligible under restricted benefits at any point during the data year.
[d] : http://cdps.ucsd.edu/index.html

designed to assess how the use of HCs relates to health care costs and utilization for vulnerable populations in the US. HCs provide comprehensive primary care and supportive services and are required to provide care for Medicaid enrollees. Recent expansions in the HC program have raised concerns about the financial sustainability of the program. Therefore, it is critical to understand if receipt of primary care in a HC has any association with health service utilization and spending for Medicaid enrollees. Of several sub-studies, we focused on a specific study using Medicaid claims data to compare Medicaid enrollees receiving primary care at HCs to non-HC users. We obtained claims data from the Medicaid Analytic eXtract (MAX) files, which is a dataset that contains individual-level protected health information. Data are not public, but available for use by researchers under a data use agreement and demonstration of adequate privacy and security protections. We defined a HC user as a patient who had more than half of their primary care visits in HCs. For this analysis, we used Illinois Medicaid claims data from 2009. We restricted the study population to adults aged 18 years and older. There were 103,379 Medicaid enrollees and roughly 19.4% of them (n = 20,061) obtained the majority of their primary care at HCs and the remaining 83,318 received care mostly elsewhere (e.g., physician office, hospital outpatients, and mixed use). The data set included not only individual-level determinants but also geographical information, Primary Care Service Area (PCSA) (Goodman *et al.*, 2003) for each beneficiary, which approximates the local geographic market for primary care. A high degree of variation across PCSAs exists; however, great homogeneity is frequently observed within a PCSA. Thus a random effect in a model is considered accounting for the clustering effect of a total of 379 PCSAs. The size of PCSA ranged from 1 to 5,655.

Due to the observational nature of the study, there are undoubtedly underlying characteristics that made patients more likely to visit HC or non-HC initially. Thus, we employed propensity scores to balance on observable characteristics between HC and non-HC. We considered potential confounding factors including patient demographics (age, sex, race), insurance characteristics (Medicaid eligibil-

Table 5: Parameter estimates and standard errors for ZIP and W-ZIP mixed models

Variable		ZIP mixed		W-ZIP^u mixed		W-ZIP^e mixed	
		Estimate	Standard error	Estimate	Standard error	Estimate	Standard error
Logistic part	Int.	-0.446^a	0.028	0.118^a	0.022	0.077^a	0.022
	HC	0.068^a	0.021	0.074^a	0.018	0.076^a	0.018
	σ_u^2	0.068	0.001	0.033	0.001	0.032	0.001
Poisson part	Int.	0.466^a	0.033	0.477^a	0.031	0.476^a	0.031
	HC	-0.034^a	0.013	-0.032^a	0.013	-0.032^a	0.013
	σ_v^2	0.171	0.001	0.152	0.001	0.154	0.001
−Log-likelihood		47640.62		47719.06		47517.61	
AIC		95293.24		95450.13		95047.23	
BIC		95315.72		95472.60		95069.70	

ZIP = zero-inflated Poisson; W-ZIP = weighted ZIP; AIC = Akaike information criteria; BIC = Bayesian information criteria.
^{a} : Indicates significance with $\alpha < 0.05$.

ity category, Temporary Aid For Needy Families (TANF) beneficiary indicator), and disease burden (childbirth, Chronic Illness and Disability Payment System for Medicaid (CDPS) risk score). Using $1:1$ nearest matching, we obtained a subsample (n = 40,122) of the dataset in which patients who did (and did not) receive their primary care service at HC were comparable. Table 4 describes the dataset before and after the propensity score matching.

Among various utilization measures of health services, we focused on the number of ED visits, which is a highly skewed distribution. Over half of individuals (68.6%) had never visited the ED over the year whereas about 0.5% (n = 564) utilized the service more than 10 days. The maximum was 133 days. We observed only 57.9% of the Medicaid patients were enrolled for the entire year and the mean observation period was 9.3 months.

The results of fitting ZIP and W-ZIP mixed models with two different weight functions are given in Table 5. The results suggest that W-ZIP^e model fits the data better than the other models in terms of smaller Akaike information criteria (AIC) and Bayesian information criteria (BIC). The estimates across all three models were very similar; HC covariate significantly affect ED visits in both logistic and Poisson parts of the models. Based on the Poisson part of the model, among patients who are potentially in need of emergency room care, HC users have less utilization than non-HC users. The health center setting may provide an efficient means of providing primary care for the Medicaid population. However, the intercept in zero-inflated part of ZIP model is less than W-ZIP models, reflecting the bias of ZIP toward underestimating utilization.

6. Discussion

In this paper, a weighted ZIP mixed model has been developed to analyze hierarchical count data with excess zeros and a variable observation time. We present two simple weight functions to handle the different individual follow-up times; however, our simulation studies show that the weighted models improve the estimate of zero-inflated part adequately. As a result, we believe our proposed model will be a useful tool to analyze properly zero-inflated count data with both random effects (clustering) and censoring in order to answer critical questions in biomedical, medical, and public health applications where such complicated data sometimes arise. To our knowledge, models that address all three issues (i.e., zero-inflation, clustering, and censoring) do not currently exist. In Medicaid health care utilization data analysis, weighted ZIP mixed model using an exponential function provides the best fit and the intercept estimate of zero-inflated part in weighted ZIP models much less bias than the unweighted ZIP model. Alternatively, different individual follow-up times can be handled by incorporating an off-

set into the Poisson part (Lee *et al.* 2001). However, we argue that a subject who experiences no event over the entire study period should be treated differently and properly from one experiencing no event but followed only partially. Model (1), therefore, incorporates a weight both the logistic and Poisson part of the model. Finally, the weight can be easily extended to a function incorporating covariates when an assumption that individual follow-up time depends on a certain covariates is reasonable. For instance, the weight function can be estimated semiparametrically based on the Cox regression model or parametrically. This represents an area of possible future research.

Appendix A:

A.1. EM algorithm

$$\begin{bmatrix}\alpha^{(k+1)}\\ u^{(k+1)}\end{bmatrix} = \begin{bmatrix}\alpha^{(k)}\\ u^{(k)}\end{bmatrix} + \Im_{\alpha,u}^{-1}\begin{bmatrix}\partial l_\xi/\partial\alpha\\ \partial l_\xi/\partial u\end{bmatrix},$$
$$\begin{bmatrix}\beta^{(k+1)}\\ v^{(k+1)}\end{bmatrix} = \begin{bmatrix}\beta^{(k)}\\ v^{(k)}\end{bmatrix} + \Im_{\beta,v}^{-1}\begin{bmatrix}\partial l_\eta/\partial\beta\\ \partial l_\eta/\partial v\end{bmatrix},$$

where

$$\Im_{\alpha,u} = \begin{bmatrix}Z'\\ K'\end{bmatrix}\left(-\frac{\partial^2 l_\xi}{\partial\xi\partial\xi'}\right)\begin{bmatrix}Z & K\end{bmatrix} + \begin{bmatrix}0 & 0\\ 0 & \sigma_u^{-2}I_m\end{bmatrix},$$
$$\Im_{\beta,v} = \begin{bmatrix}X'\\ K'\end{bmatrix}\left(-\frac{\partial^2 l_\eta}{\partial\eta\partial\eta'}\right)\begin{bmatrix}X & K\end{bmatrix} + \begin{bmatrix}0 & 0\\ 0 & \sigma_v^{-2}I_m\end{bmatrix},$$
$$\frac{\partial l_\xi}{\partial\alpha} = Z'\frac{\partial l_\xi}{\partial\xi}, \qquad \frac{\partial l_\xi}{\partial u} = K'\frac{\partial l_\xi}{\partial\xi} - \frac{u}{\sigma_u^2},$$
$$\frac{\partial l_\eta}{\partial\beta} = X'\frac{\partial l_\eta}{\partial\eta}, \qquad \frac{\partial l_\xi}{\partial v} = K'\frac{\partial l_\eta}{\partial\eta} - \frac{v}{\sigma_v^2}$$

and

$$\frac{\partial l_\xi}{\partial\xi_{ij}} = 1 - \frac{\psi_{ij}}{1+\exp(\xi_{ij}) - w_{ij}\exp(\xi_{ij})} - \frac{\exp(\xi_{ij})}{1+\exp(\xi_{ij})},$$
$$\frac{\partial l_\eta}{\partial\eta_{ij}} = (1-\psi_{ij})\left(y_{ij} - \exp(\eta_{ij})\right).$$

Then we have

$$\frac{\partial l_\xi}{\partial\xi} = \frac{\psi}{1+\exp(\xi) - w\exp(\xi)} - \frac{\exp(\xi)}{1+\exp(\xi)},$$
$$\frac{\partial l_\eta}{\partial\eta} = (1-\psi)(y - \exp(\eta))$$

and

$$-\frac{\partial^2 l_\xi}{\partial\xi\partial\xi'} = \mathrm{Diag}\left[\frac{\psi\exp(\xi)(w-1)}{(1+\exp(\xi) - w\exp(\xi))^2} + \frac{\exp(\xi)}{(1+\exp(\xi))^2}\right],$$
$$-\frac{\partial^2 l_\eta}{\partial\eta\partial\eta'} = \mathrm{Diag}\left[(1-\psi)\exp(\eta)\right].$$

A.2. Variance component estimation

Suppose $\Im_{\alpha,u}$ is partitioned conformally to :

$$\alpha|u \text{ as } \Im^{-1}_{\alpha,u} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}.$$

Suppose $\Im_{\beta,v}$ is partitioned conformally to :

$$\beta|v \text{ as } \Im^{-1}_{\beta,v} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}.$$

Then estimators of variance components are given by:

$$\hat{\sigma}^2_u = \frac{1}{m}\left(\text{tr}(A_{22}) + \hat{u}'\hat{u}\right),$$
$$\hat{\sigma}^2_v = \frac{1}{m}\left(\text{tr}(B_{22}) + \hat{v}'\hat{v}\right).$$

In addition, the asymptotic variance matrix of the variance component estimators are obtained from the inverse of the residual maximum likelihood information matrix (McGilchrist and Yau, 1995) as follows:

$$\text{var}\begin{bmatrix} \hat{\sigma}^2_u \\ \hat{\sigma}^2_v \end{bmatrix} = 2\begin{bmatrix} \sigma_u^{-4}\text{tr}\left(I_m - \sigma_u^{-2}A_{22}\right)^2 & 0 \\ 0 & \sigma_v^{-4}\text{tr}\left(I_m - \sigma_v^{-2}B_{22}\right)^2 \end{bmatrix}^{-1}.$$

Appendix B:

R for fitting weighted ZIP models.

```
[language=R]
# m = # clusters, ni=# subjects within cluster i, N=sum of ni
# Y(Nx1) outcomes
# Z(pz x N) covariates for logit part
# X(px x N) covariates for Poisson part
# K=cluster indicator (mxN)
# w(Nx1) weight
# alpha, beta, sig2u, sig2v : initial values
library(psych)
estpm =function(Y,Z,X,K,w, alpha=NULL, beta=NULL, sig2u=NULL, sig2v=NULL){
  N = dim(Y)[1]
  pz = dim(Z)[1]
  px = dim(X)[1]
  m = dim(K)[1]
  pzm=pz+m
  pxm=px+m
  if (is.null(alpha)){alpha=as.matrix(c(1,1))}
  if (is.null(beta)){beta=as.matrix(c(1,2))}
  if (is.null(sig2u)){sig2u=0.5}
```

```
if (is.null(sig2v)){sig2v=0.5}
u=as.matrix(rnorm(m, sd=sqrt(sig2u)),m,1)
v=as.matrix(rnorm(m, sd=sqrt(sig2v)),m,1)

it = 0; dta = 1
itMax=1e4; eps=1e-5
while( (it<itMax)&(dta>eps) ){
  ex_xi=exp(t(Z)%*%alpha+t(K)%*%u)
  ex_eta=exp(t(X)%*%beta+t(K)%*%v)
  pi=ex_xi/(1+ex_xi)

  #1) psi
  psi=(1-w*pi)/(1-w*pi+w*pi*exp(-ex_eta))
  psi[Y>0]=0

  #2) first derivatives
  d1=1-psi/(1+ex_xi-w*ex_xi)-ex_xi/(1+ex_xi)
  d2=(1-psi)*(Y-ex_eta)
  dalpha=Z%*%d1 #pzx1
  dbeta=X%*%d2 #pxx1
  du=K%*%d1-u/sig2u #mx1
  dv=K%*%d2-v/sig2v

  #3) "-2nd derivatives"
  dd1=-psi*ex_xi*(1-w)/(1+ex_xi-w*ex_xi)^2+ex_xi/(1+ex_xi)^2
  dd2=(1-psi)*ex_eta

  #4) inversion matrix
  tmp1=T1=matrix(0,pzm,pzm) ; tmp2=T2=matrix(0,pxm,pxm)
  ZK=rbind(Z,K)
  XK=rbind(X,K)
  diag(tmp1[(pz+1):pzm,(pz+1):pzm])=1/sig2u
  diag(tmp2[(px+1):pzm,(px+1):pzm])=1/sig2v
  for (i in 1:pzm){for (j in 1:pzm){T1[i,j]=sum(ZK[i,]*ZK[j,]*dd1)} }
  for (i in 1:pxm){for (j in 1:pxm){T2[i,j]=sum(XK[i,]*XK[j,]*dd2)} }
  T1=T1+tmp1
  T2=T2+tmp2

  A1=T1[1:pz,1:pz]
  B1=T1[1:pz,(pz+1):(pz+m)]
  C1=T1[(pz+1):(pz+m),1:pz]
  D1=T1[(pz+1):(pz+m),(pz+1):(pz+m)]
  A2=T2[1:px,1:px]
  B2=T2[1:px,(px+1):(px+m)]
  C2=T2[(px+1):(px+m),1:px]
  D2=T2[(px+1):(px+m),(px+1):(px+m)]
```

```
    ai1=solve(A1)
    di1=solve(D1)
    ai2=solve(A2)
    di2=solve(D2)

    AI1=solve(A1-B1%*%di1%*%C1)
    DI1=solve(D1-C1%*%ai1%*%B1)
    AI2=solve(A2-B2%*%di2%*%C2)
    DI2=solve(D2-C2%*%ai2%*%B2)
    BI1=-ai1%*%B1%*%DI1
    CI1=-di1%*%C1%*%AI1
    BI2=-ai2%*%B2%*%DI2
    CI2=-di2%*%C2%*%AI2

    # 5) update
    Alpha=alpha+AI1%*%dalpha+BI1%*%du
    Beta=beta+AI2%*%dbeta+BI2%*%dv
    U=u+CI1%*%dalpha+DI1%*%du
    V=v+CI2%*%dbeta+DI2%*%dv
    Sig2u=(tr(DI1)+t(u)%*%u)/m
    Sig2v=(tr(DI2)+t(v)%*%v)/m

    dta = mean(c(abs(Alpha-alpha),abs(U-u),abs(Beta-beta),abs(V-v)
                 ,abs(Sig2u-sig2u),abs(Sig2v-sig2v)))
    alpha = Alpha
    u = U
    beta = Beta
    v = V
    sig2u = as.vector(Sig2u)
    sig2v = as.vector(Sig2v)
    it = it + 1
  }
  pi=ex_xi/(1+ex_xi)
  lambda=ex_eta
  loglik=sum((log(1-w*pi+w*pi*exp(-lambda)))[Y==0])
         +sum((log(w)+log(pi)+dpois(Y,lambda,log=TRUE))[Y>0])
         +sum(dnorm(u,sd=sqrt(sig2u),log=TRUE)+dnorm(v,sd=sqrt(sig2v),
              log=TRUE))
  XM=matrix(0,m,m);diag(XM)=1
 a.comp=(tr(XM-DI1/sig2u)/sig2u)^2
 b.comp=(tr(XM-DI2/sig2v)/sig2v)^2
 var_sig2u=2/a.comp
 var_sig2v=2/b.comp
 out=list(alpha=alpha, beta=beta, u=u,v=v, sig2u=sig2u, sig2v=sig2v, var1=AI1
           , var2=AI2, loglik=loglik,var_sig2u=var_sig2u,var_sig2v=var_sig2v)
}
```

Acknowledgements

This work was funded under a contract to the Health Resources and Services Administration (HRSA). Medicaid claims data used in the application section is not available under the contract.

References

Emerson SS, McGee DL, Fennerty B, Hixson L, Garewal H, and Alberts D (1993). Design and analysis of studies to reduce the incidence of colon polyps, *Statistics in Medicine*, **12**, 339–351.

Goodman DC, Mick SS, Bott D, *et al.* (2003). Primary care service areas: a new tool for the evaluation of primary care services, *Health Services Research*, **38**, 287–309.

Hall DB (2000). Zero-inflated and binomial regression with random effects: a case study, *Biometrics*, **56**, 1030–1039.

Hsu CH (2005). Joint modelling of recurrence and progression of adenomas: a latent variable approach, *Statistical Modelling*, **5**, 201–215.

Hsu CH (2007). A weighted zero-inflated Poisson model for estimation of recurrence of adenomas, *Statistical Methods in Medical Research*, **16**, 155–166.

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

Lee AH, Wang K, Scott JA, Yau KK, and McLachlan GJ (2006). Multi-level zero-inflated Poisson regression modelling of correlated count data with excess zeros, *Statistical Methods in Medical Research*, **15**, 47–61.

Lee AH, Wang K, and Yau KK (2001). Analysis of zero-inflated Poisson incorporating extent of exposure, *Biometrical Journal*, **43**, 963–975.

McGilchrist C and Yau K (1995). The derivation of BLUP, ML, REML estimation methods for generalised linear mixed models, *Communications in Statistics-Theory and Methods*, **24**, 2963–2980.

Min Y and Agresti A (2005). Random effect models for repeated measures of zero-inflated count data, *Statistical Modelling*, **5**, 1–19.

Moghimbeigi A, Eshraghian MR, Mohammad K, and McArdle B (2008). Multilevel zero-inflated negative binomial regression modeling for over-dispersed count data with extra zeros, *Journal of Applied Statistics*, **35**, 1193–1202.

Monod A (2014). Random effects modeling and the zero-inflated Poisson distribution, *Communications in Statistics*, **43**, 664–680.

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

Ridout M, Demétrio CG, and Hinde J (1998). Models for count data with many zeros. In *Proceedings of the XIXth International Biometric Conference*, **19**, 179–192.

Yau KK and Lee AH (2001). Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme, *Statistics in Medicine*, **20**, 2907–2920.

Received September 28, 2017; Revised December 23, 2017; Accepted December 24, 2017