調查研究—方法與應用/第 47 期
2021 年 10 月，頁 95-111

Research Articles

Modeling Multiple Inflations in Survey Data

Ting Hsiang Lin*, Min-Hsiao Tsai**

ABSTRACT

The Poisson (ZIP) regression model is used to analyze data with a Poisson distribution with excessive zeros. Although various models have been developed to fit zero-inflated data, many of them strongly depend on unique features of each data set. To be more specific, this means a sizable group of respondents endorsing the same answers, making the data have modes. For example, some data have cyclical patterns with multiple inflated values, such as survey questions which assess risk or health behaviors within a fixed length of time. Two examples are the question "During the past 30 days, on how many days did symptoms of asthma make it difficult for you to stay asleep?" and the question "During the past two weeks, on how many days did you text or e-mail while driving a car or other vehicle?"

In this study, we proposed a new multiple-inflated truncated Poisson (MITP) regression model for more than two inflated values. The model is a combination of multinomial logistic regression and

* Corresponding Author. Professor, Department of Statistics, National Taipei University. 151, University Rd., San Shia District, New Taipei City, 23741 Taiwan. 886-2-86741111 ext. 66767. E-mail: tinghlin@mail.ntpu.edu.tw.

** Professor, Department of Statistics, National Taipei University.

truncated Poisson regression; the multinomial logistic regression models the occurrence of excessive values, and the truncated Poisson regression models data following a truncated Poisson distribution. The performance of the proposed model was evaluated through a simulation study. In the simulation study, we compared the performance of truncated Poisson (TP), zero-inflated truncated Poisson (ZITP), zero- and K-inflated Poisson (ZKIP) and multiple-inflated truncated Poisson (MITP) regression models under different simulation configurations. The factors considered include the model used to generate the data and the sample size. A likelihood ratio test was used to select the best model. We generated 1000 replications for each configuration. The accuracy rates of model selection via the likelihood ratio test and MAE (mean absolute error) were used to compare the performance of the models. In terms of MAE, when the hypothetical true model is TP, the means of the MAE of the four models do not have any substantial difference. When the hypothetical true model is ZITP, TP has the worst performance. When the hypothetical true model is ZKITP, TP and ZITP perform poorly, whereas the performance of ZKITP and MITP is much better. When the hypothetical true model is MITP, MITP has the best performance. From the results, MITP is the best model when there are multiple inflated points. ZKITP and MITP fit well when there are zero and K inflated points, while ZITP, ZKITP and MITP fit the data well when there are only inflated zero counts. When the data are truncated Poisson distributed, all four models fit the data well. With an increasing K-inflation rate, ZKITP and MITP have better and stable performance. With fixed sample sizes and parameters, when the true underlying model is truncated Poisson, MITP has the smallest MAE, followed by ZKITP, ZITP and TP.

We analyzed a survey question, “On how many of the PAST 30 DAYS did you smoke cigarettes?” from the US’s National Adult Tobacco Survey (NATS) in the empirical study. In addition to typical days of the Poisson distribution (1, 2, 3, and 4 days), the data also

have inflated values that are multiples of 5 and 7. The results indicate that the MITP model has smaller MAE, and MSE as well as better model fit, and outperformed competing models.

Keywords: truncated Poisson regression, multinomial logistic regression, zero-inflated Poisson

處理具有多個膨脹值之問卷調查資料摘要

林定香* 蔡旻曉**

摘要

零膨脹卜瓦松（ZIP）迴歸模型主要用於分析有過多零值的資料。雖然已有許多模型用來處理膨脹資料，但多數模型仍須高度仰賴於資料的獨特性。一般而言，當受訪者回答了相同的答案，使得資料出現了一些峰值，就產生了膨脹值資料。本文中，我們提出了一個新的多點膨脹截斷卜瓦松迴歸模型（MITP），可以對多個膨脹值進行建模，該模型是多項式邏輯斯模型和截斷卜瓦松迴歸的混合模型，其中多項式邏輯斯模型預測膨脹值的發生與否。截斷卜瓦松迴歸對呈現截斷卜瓦松分配的計數資料進行建模。在實證研究中，我們以國家成人煙草調查（NATS）中的一個問題“您過去 30 天內有多少天吸煙”為例，資料除了典型的卜瓦松分配天數外（1、2、3、4 天等），在 5 天及 7 天的倍數天數也有明顯的膨脹值。結果顯示我們的模型比其他競爭模型具有更佳的配適度。

關鍵詞：截斷卜瓦松迴歸，多項式邏輯斯迴歸，零膨脹卜瓦松

* 通訊作者，國立臺北大學統計學系教授。
** 國立臺北大學統計學系教授。

1. Introduction

Zero-inflated data are prevalent in a wide variety of subjects, and the zero-inflated Poisson (ZIP) regression model (Lambert 1992) and its variants are commonly used to analyze data with a mass of zeros. There are two ways to handle excessive zero values, and they can be treated as a mixture model or a two-component model. When the model is regarded as a mixture model, the data can be fitted by a standard discrete distribution with inflated zeros, or a semi-continuous distribution with positive continuous values combined with a substantial portion of zeros. When the model is considered a two- component model, one component handles zeros and the other component handles positive values. The zero component is fitted by a logistic regression model for the occurrence of an event, and the non-zero component can be modelled by a Poisson or a log-normal regression for the positive values that are conditional on the occurrence of the event (Welsh et al. 1996; Zhou and Tu 1999).

The problem with ZIP is that it only models one single inflated data value, and the inflated value must be zero. Models with inflated values other than zero have been proposed by several researchers (Famoye and Singh 2003; Bae et al. 2005), who considered data with another massive value *K*. The ZIP has been extended to a zero- and *K*-inflated Poisson (ZKIP) regression model (Lin and Tsai 2013; Tsai and Lin 2017) that models data with two inflated values concurrently. The model is a combination of multinomial logistic regression and Poisson regression. The multinomial logistic regression models the occurrence of excessive values, including zero and *K*.

In a real data situation, some data have cyclical patterns with multiple

inflated values. The preponderance of some particular values occurs due to the nature of the data. For instance, survey questions may measure the frequency of human behaviors within a fixed time frame, such as survey questions assessing risk or health behaviors during a length of time, for example, the question "During the past 30 days, on how many days did symptoms of asthma make it difficult for you to stay asleep?" in the 2014 Behavioral Risk Factor Surveillance System (BRFSS), and the question "During the past 30 days, on how many days did you text or e-mail while driving a car or other vehicle?" in the 2015 Youth Risk Behavior Survey (YRBS. Centers for Disease Control and Prevention 2014; 2015).

One typical example is "On average, on how many days during the past 30 days did you drink alcohol?" For such a question, in addition to the responses of zero and thirty days for non-drinkers and daily drinkers respectively, there would be some inflated values of multiples of 5 or 7 days, including 10, 14, 15, 20, etc. (Wang and Heitjan 2008). Another example is "On average, how many cigarettes do\ did you smoke each day?" In addition to responses of zero for non-smokers, there would be some inflated values of multiples of 5 and 10. There are two possible causes for respondents to answer with multiples of 5 or 7. First, the subjects may not recall precisely the number of days on which they have drunk during the past 30 days, and thus report a rounded approximation. Second, the subjects have a tendency to provide an approximation with a natural unit, such as "a week" (7 days) or "a pack" (20 cigarettes). Thus, we would encounter data with multiple inflated values.

In this study, we propose a model to handle data with multiple inflated values. The proposed model is a combination of multinomial logistic regression and truncated Poisson regression. The multinomial logistic regression models the occurrence of excessive values. The truncated Poisson regression

models the values that follow a truncated Poisson distribution. The remainder of the article is organized as follows. Section 2 formulates our new multiple inflated truncated Poisson regression model. In Section 3, we present a simulation study. In Section 4, we analyze empirical data with our proposed model and compare it with some competing models. Finally, we conclude the study with discussion in section 4.

2. Model formulation

When there are multiple inflated values, let's say, R of them, and the data take integer values between 0 and K, we can define the model as follows:

$$\underset{i=1,\ldots,n}{\overset{\text{indep.}}{Y_i}} \sim P(Y_i=y_i|\mu_i,\psi_{i,1},\ldots,\psi_{i,R})=\begin{cases}\psi_{i,j}+\pi_i\left(\frac{\mu_i^{y_i}}{y_i!}\right)\left(\sum_{z=0}^{K}\frac{\mu_i^z}{z!}\right)^{-1}, & y_i=a_j\in\{0,1,\ldots,K\}, j=1,\ldots,R \\ \pi_i\left(\frac{\mu_i^{y_i}}{y_i!}\right)\left(\sum_{z=0}^{K}\frac{\mu_i^z}{z!}\right)^{-1}, & y_i\in\bar{A}=\{0,1,\ldots,K\}/\{a_j\}_{j=1}^{R}\end{cases}$$

$$=\pi_i\times\prod_{j=1}^{R}\left[\frac{\psi_{i,j}}{\pi_i}+\left(\frac{\mu_i^{y_i}}{y_i!}\right)\left(\sum_{z=0}^{K}\frac{\mu_i^z}{z!}\right)^{-1}\right]^{I_{\{a_j\}}(y_i)}\times\left[\left(\frac{\mu_i^{y_i}}{y_i!}\right)\left(\sum_{z=0}^{K}\frac{\mu_i^z}{z!}\right)^{-1}\right]^{I_{\bar{A}}(y_i)},$$

where K is the truncated value, $\psi_{i,j}, j=1,\ldots,R$ represents the proportions of inflated values that do not follow a truncated Poisson distribution, and $\pi_i=1-\sum_{j=1}^{R}\psi_{i,j}$ represents the proportion of values that belong to the true underlying truncated Poisson model. All $\psi_{i,j}$ and π_i are in the interval [0,1]. Those $I_{\{aj\}}(y_i)$ are indicator variables for occurrence of inflated values, and $I_{\{aj\}}(y_i)=1$ if $y_i=a_j$, and 0 otherwise; $I_{\bar{A}}(y_i)=1$ if y_i is not an inflated value.

Note that when $R=2$, the MITP model is simplified to a zero- and K-inflated truncated Poisson (ZKITP) regression model, i.e.,

$$\underset{i=1,\ldots,n}{\overset{\text{indep.}}{Y_i \sim P(Y_i=y_i|\mu_i,\psi_{i,1},\psi_{i,2})}}=\begin{cases}\psi_{i,j}+\pi_i\left(\dfrac{\mu_i^{y_i}}{y_i!}\right)\left(\sum\limits_{z=0}^{K}\dfrac{\mu_i^z}{z!}\right)^{-1}, & y_i=a_j\in\{0,1,\ldots,K\}, j=1,2\\ \pi_i\left(\dfrac{\mu_i^{y_i}}{y_i!}\right)\left(\sum\limits_{z=0}^{K}\dfrac{\mu_i^z}{z!}\right)^{-1}, & y_i\in\bar{A}=\{0,1,\ldots,K\}/\{a_j\}_{j=1}^2\end{cases}$$

When $R=1$ the MITP model is simplified to a zero-inflated truncated Poisson (ZITP) regression model, i.e.,

$$\underset{i=1,\ldots,n}{\overset{\text{indep.}}{Y_i \sim P(Y_i=y_i|\mu_i,\psi_i)}}=\begin{cases}\psi_i+\pi_i\left(\dfrac{\mu_i^{y_i}}{y_i!}\right)\left(\sum\limits_{z=0}^{K}\dfrac{\mu_i^z}{z!}\right)^{-1} & y_i=0\\ \pi_i\left(\dfrac{\mu_i^{y_i}}{y_i!}\right)\left(\sum\limits_{z=0}^{K}\dfrac{\mu_i^z}{z!}\right)^{-1} & y_i\in\bar{A}=\{0,1,\ldots,K\}/\{0\}\end{cases}$$

and when R is not specified, the model is simplified, and reduced to a truncated Poisson (TP) regression model, i.e.,

$$\underset{i=1,\ldots,n}{\overset{\text{indep.}}{Y_i \sim P(Y_i=y_i|\mu_i)}}=\left(\frac{\mu_i^{y_i}}{y_i!}\right)\left(\sum_{z=0}^{K}\frac{\mu_i^z}{z!}\right)^{-1}, y_i=a_j\in\{0,1,\ldots,K\},$$

3. Simulation study

A simulation study is conducted to evaluate the proposed model. We generated data for a truncated Poisson (TP) regression model, zero-inflated truncated Poisson (ZITP) regression model, zero- and K-inflated truncated Poisson (ZKITP) regression model, and multiple inflated truncated Poisson (MITP) regression model.

In this simulation, we set K at 14 (two weeks), and one covariate x_i follows a normal distribution. Since the expected values will vary with x_i, β_0 and β_1, we set up $\beta_0=-1.4$ and $\beta_1=0.1$, and generated x_i data within 3 standard deviations to ensure the generated data are less or equal to 14. By plugging in

the equation between μ_i and x_i, the range of μ_i is within 0.6703 to 13.4637. The values of $(\psi_1,\psi_2,\psi_3,\psi_4,\psi_5)$ are set as (0, 0, 0, 0, 0), (0.25, 0, 0, 0, 0), (0.25, 0.20, 0, 0, 0), and (0.25, 0.20, 0.10, 0.05, 0.10) for the TP, ZITP, ZKITP, and MITP models, respectively. The sample sizes n considered here are 100 and 250. We generated 1000 replications for each configuration. The MAE (mean absolute error) is used to evaluate the models' performance:

$$\mathrm{MAE}=\frac{1}{m}\sum_{y}|\mathrm{O}_y-\mathrm{E}_y|$$

Table 1 displays the simulated results of MAE and accuracy rate of model selection when the sample size n is 100. When the hypothetical true model is TP, the means of the MAE of the four models do not have substantial difference. When the hypothetical true model is ZITP, TP has the worst performance, the mean MAE of TP is about 2.68 times the mean MAE of ZITP, and the mean MAE are similar for the other three models. When the hypothetical true model is ZKITP, TP and ZITP perform poorly and the mean MAE of TP and ZITP are about 6.49 and 4.88 times the mean MAE of ZKITP. The performance of ZKITP and MITP is much better, and the mean MAE are similar for both models. When the hypothetical true model is MITP, MITP has the best performance, and the mean MAE of TP, ZITP, and ZKITP are about 10.53, 7.53 and 3.69 times the mean MAE of MITP, respectively.

The accuracy rates of model selection via likelihood ratio test are also presented in Table 1. When the data are generated based on the TP model, 97.2% are fitted as TP, 1.6% are fitted as ZITP, 0.5% are fitted as ZKITP, and 0.7% are fitted as MITP. When the data are generated based on the ZITP model, 98.9% are fitted as ZITP, 0.1% are fitted as ZKITP, and 1.0% are fitted as MITP. When the data are generated based on the ZKITP model,

Table 1. The MAE and accuracy rate of model selection when sample size n = 100

True Model	Fitted Model			
	TP	ZITP	ZKITP	MITP
TP	1.5566 97.2%	1.5366 1.6%	1.5345 0.5%	1.4711 0.7%
ZITP	3.3863 0.0%	1.2637 98.9%	1.2628 0.1%	1.2075 1.0%
ZKITP	6.9976 0.0%	5.2571 0.0%	1.0783 99.5%	1.0305 0.5%
MITP	7.0992 0.0%	5.0760 0.0%	2.4882 0.0%	0.6741 100.0%

For each configuration, the entries in the first row are MAEs, and the entries in the second row are accuracy rates of the fitted models.

99.5% are fitted as ZKITP and 0.5% are fitted as MITP. When the data are generated based on the MITP model, 100% are fitted as MITP.

Table 2 shows the simulated results of MAE and the accuracy rate of model selection when the sample size increases to 250. When the hypothetical true model is TP, the means of the MAE of the four models do not differ substantially. When the hypothetical true model is ZITP, TP has the worst performance, the mean MAE of TP is about 3.91 times the mean MAE of ZITP, and the mean MAE are similar for the other three models. When the hypothetical true model is ZKITP, TP and ZITP perform poorly and the mean MAE of TP and ZITP are about 10.09 and 7.68 times the mean MAE of ZKITP. The performance of ZKITP and MITP is much better, and the mean MAE is similar for both models. When the hypothetical true model is MITP, MITP has the best performance, and the mean MAE of TP, ZITP, and ZKITP are about 16.59, 11.84 and 5.77 times the mean MAE of MITP, respectively.

Table 2. The MAE and accuracy rate of model selection when sample size n=250

True Model	Fitted Model			
	TP	ZITP	ZKITP	MITP
TP	2.4663 96.5%	2.4405 1.8%	2.4389 0.6%	2.3762 1.1%
ZITP	7.9421 0.0%	2.0296 99.1%	2.0282 0.2%	1.9531 0.7%
ZKITP	17.3120 0.0%	13.1861 0.0%	1.7163 98.9%	1.6478 1.1%
MITP	17.4489 0.0%	12.4527 0.0%	6.0653 0.0%	1.0520 100.0%

For each configuration, the entries in the first row are MAEs, and the entries in the second row are accuracy rates of the fitted models.

Moreover, when the data are generated based on the TP model, 96.5% are fitted as TP, 1.8% are fitted as ZITP, 0.6% are fitted as ZKITP, and 1.1% are fitted as MITP. When the data are generated based on the ZITP model, 99.1% are fitted as ZITP, 0.2% are fitted as ZKITP, and 0.7% are fitted as MITP. When the data are generated based on the ZKITP model, 98.9% are fitted as ZKITP and 1.1% are fitted as MITP. When the data are generated based on the MITP model, 100% are fitted as MITP.

4. An empirical application

We now compare the fit of the MITP model to the other models for data from the Us's National Adult Tobacco Survey (NATS). The NATS was created to measure how widespread tobacco use is among adults, as well as what factors lead to or prevent such use. The NATS also establishes a frame-

work for evaluating the national as well as state-specific tobacco control programs. The NATS used a stratified, national, landline and cell phone survey of non-institutionalized adults aged 18 years and older residing in the 50 states or Washington D.C. area. The survey was developed to collect representative and comparable data at both national and state levels. One of the aims of the sample design is to provide national estimates for subgroups stratified by gender, age, and ethnicity. The NATS is designed within the framework defined by the Office of Smoking and Health's Key Outcome Indicators (KOI) report. The NATS questionnaire is developed around KOI to achieve the following goals: to prevent initiation of tobacco use; to eliminate nonsmokers' exposure to second-hand smoke, to promote quitting and to eliminate tobacco-related disparities. The response variable is "On how many of the PAST 30DAYS did you smoke cigarettes?" The covariates used in this analysis are age in years and self-evaluated health status (Would you say that in general your health is…?) and the five options are "excellent", "very good", "good", "fair" and "poor". The data were fitted to truncated Poisson (TP) regression, zero-inflated truncated Poisson (ZITP) regression, zero- and K-inflated truncated Poisson (ZKITP) regression, and multiple inflated truncated Poisson (MITP) regression models, with age and self-reported health status of the respondents as predictors. All the parameter estimates with their standard errors in are obtained by the NLM (non-linear minimization) function in software R. The NLM function applies a Newton-type algorithm and implements a minimization of a negative log-likelihood function with extra analytical gradients. The standard errors are derived with the Hessian matrix returned numerically by NLM. The log-likelihood of the MITP model is larger than that of the other three models. The log-likelihood of MITP, ZKITP, ZITP and TP is −11,356.89, −13,555.05, −15,189.93

and −65,307.345. Since the models are nested, a likelihood ratio test can be performed for model comparison with chi-square statistics, and the difference of the twice of negative log-likelihood follows a chi-square distribution. The results show MITP outperforms ZKITP, ZKITP outperforms ZITP, and ZITP outperforms TP. We also use information criteria as a model selection index, AIC and BIC. As shown in Table 3, the values of AIC are 23,519.78, 27,358.1, 30,565.86, and 130,676.69, and the values of BIC are 22,723.67, 27,119.99, 30,389.75, and 130,624.60 for MITP, ZKITP, ZITP and TP, respectively. Table 3 indicates that MTIP has the smallest AIC and BIC, which suggests that MTIP is the best model of the four.

Table 3. Observed frequency and residuals of each fitted models.

	Value	Observed Freq	MITP	ZKITP	ZITP	TP
Inflated	0	17,874	0.00	0.00	0.00	10,293.60
	5	113	−4.08	97.89	108.27	−362.27
	7	39	−61.34	−16.67	16.03	−133.17
	10	177	−2.59	26.97	81.31	138.63
	14	27	−1.04	−126.47	−155.75	22.42
	15	282	−0.18	153.54	101.64	279.35
	20	304	0.00	282.08	229.78	303.82
	21	9	0.00	−4.20	−44.92	8.89
	25	89	0.00	87.88	78.86	88.99
	28	22	0.00	21.88	19.99	22.00
	30	245	0.00	−0.02	244.41	245.00
Non-inflated	1	92	85.67	91.92	91.99	−5512.97
	2	107	86.70	106.53	106.90	−2,864.90
	3	82	38.20	80.04	81.53	−1,442.88

Table 3. Observed frequency and residuals of each fitted model (continued)

Value	Observed Freq	MITP	ZKITP	ZITP	TP
4	62	−9.52	55.92	60.34	−764.60
6	31	−73.58	−0.31	19.76	−252.63
8	23	−62.03	−63.70	−18.25	−81.92
9	3	−61.64	−117.16	−63.09	−60.71
11	3	−25.27	−167.49	−123.47	−19.88
12	32	15.44	−145.78	−121.90	18.51
13	2	−7.03	−169.32	−171.64	−5.88
16	4	2.99	−96.92	−163.68	2.46
17	11	10.56	−63.71	−136.45	10.10
18	13	12.82	−39.29	−110.06	12.47
19	4	3.93	−30.72	−93.80	3.69
22	9	9.00	1.40	−28.58	8.94
23	5	5.00	0.81	−20.19	4.96
24	9	9.00	6.79	−7.27	8.98
26	11	11.00	10.45	4.90	10.99
27	10	10.00	9.74	6.44	10.00
29	8	8.00	7.95	6.90	8.00
Log-likelihood		−11,356.89	−13,555.05	−15,189.93	−65,307.345
MAE		19.89	67.21	81.23	742.05
MSE		1,127.49	9,009.33	10,855.24	4,764,788.44
AIC		23,519.78	27,358.10	30,565.86	130,676.69
BIC		22,723.67	27,119.99	30,389.75	130,624.60

Using the probability density function of the aforementioned distributions and the samples size n, we calculate the predicted frequencies for each day. The predicted frequencies can be obtained as follows:

$$E_y = nP(Y=y) = \sum_{i=1}^{n} P(Y=y \mid X=x_i),\ y=0,1,2,...,K$$

where X are the covariates. To measure the overall accuracy of the proposed parameter's estimation, we adopt MAE and MSE (mean square error)

$$\text{MSE} = \frac{1}{m}\sum_{y}(\text{O}_y - \text{E}_y)^2$$

where O_y are the observed frequencies, E_y are the predicted frequencies, and m is the total number of different values, which is 31 in our example. Both MAE and MSE are smallest for MITP and biggest for TP, suggesting that MITP is best at making predictions among all the models and TP is the worst.

Table 3 displays the observed frequencies and residuals. The top panel is for inflated values, and MITP has the smallest residuals, followed by ZKITP, then ZITP, and TP has the largest ones. The three models MITP, ZKITP and ZITP have a perfect fit for a value of zero, and similarly, MITP and ZKITP fit the inflated value "30" quite well. These results are expected because the value "zero' is imposed into MITP, ZKITP and ZITP, and similarly, "30" is imposed into MITP and ZKITP. The MITP fits better for all inflated values than the other models except for the value 7. The bottom panel is for the non-inflated values. When predicting the values under 5, MITP performs substantially better than the other models. ZKITP and ZITP are equally good, while TP fits poorly. In general, MITP has more positive residuals, which implies MITP tends to underpredict the values, while TP tends to overpredict. There is no particular pattern for ZKITP and ZITP. The residuals are smallest for MITP for most values. For some values, the predictions of MITP are not as good as the other models and they are

roughly adjacent values of the inflated values. For example, the residuals of the values 6 and 8 of MITP are not necessarily smallest among the four models.

5. Conclusion

In this study, we extended ZIP for a truncated Poisson distribution and proposed a new multiple inflated truncated Poisson regression model for fitting data with multiple inflated values. The model has several variations under different conditions. The first extension is the zero inflated truncated Poisson (ZITP) with inflated zeros; the second extension is the zero- and *K*-inflated truncated Poisson (ZKITP) with two inflated values. Lastly, multiple inflated truncated Poisson (MITP) is a truncated Poisson regression with several (more than two) inflated values. All these models can be considered mixture models of two components. The first component of the models is to fit the occurrence of the inflated values, and it is fitted by a binary logistic model for ZITP, and a multinomial logistic model for ZKITP and MITP. The second component of the models handles the non-inflated values, and is fitted by a truncated Poisson regression model. The major difference in the current study is that we have extended dual inflated values to multiple inflated values. When there are dual inflated values and they are endpoints of the data, it usually represents completely non-committed or completely committed behaviors. For example, when addressing drinking behavior in 30 days, 0 and 30 days represent non-drinker and daily drinker. In this study, we extended to with multiple inflation values in addition to dual endpoints, and further explore the behavior of multiples of 5 or 7 days.

In the simulation study, we evaluated the performance of the four mod-

els by MAE under different configurations, including sample size and inflation rates of inflated points. From the result, MITP is the best model when there are multiple inflated points. ZKITP and MITP fit well when there are zero and *K* inflated points, and ZITP, ZKITP and MITP fit the data well when there are only inflated zero counts. When the data are truncated Poisson distributed, all four models fit the data well. With an increasing *K*-inflation rate, ZKITP and MITP have better and stable performance. With fixed sample sizes and parameters, when the true underlying model is truncated Poisson, MITP has the smallest MAE, followed by ZKITP, ZITP and TP.

We used the National Adult Tobacco Survey (NATS) to compare the performance of the four models as an empirical study. We found the multiple inflated truncated Poisson (MITP) regression model outperforms other models in both model selection and prediction of values'. The MITP is especially precise in predicting inflated values, while ZKITP fits well with 0 and 30 and ZITP fits well with 0.

In conclusion, the MITP is a reliable option when analyzing repetitive events in a fixed length of time with multiple massive values, and the model we have proposed is sufficiently flexible to incorporate all types of independent variables with different properties, including both discrete and continuous quantitative variables as well as qualitative variables such as those in generalized linear models

REFERENCES

Bae, Sejong, Felix Famoye, John T. Wulu, Alfred A. Bartolucci, and Karan P. Singh, 2005, "A Rich Family of Generalized Poisson Regression Models with Applications." *Mathematics and Computers in Simulation* 69(1-2): 4-11.

Centers for Disease Control and Prevention, 2014, "Behavioral Risk Factor Surveillance

System Survey Questionnaire." Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.

——, 2015, "Youth Risk Behavior Survey Questionnaire." http://www.cdc.gov/yrbs (Date visited: May 30, 2021).

Famoye, Felix, and Karan P. Singh, 2003, "On Inflated Generalized Poisson Regression Models." *Advances and Applications in Statistics* 3(2): 145-158.

Lambert, Diane, 1992, "Zero-inflated Poisson Regression, with an Application to Defects in Manufacturing." *Technometrics* 34(1): 1-14.

Lin, Ting Hsiang, and Min-Hsiao Tsai, 2013, "Modeling Health Survey Data with Excessive Zero and K Responses." *Statistics in Medicine* 32(9): 1572-1583.

Tsai, Min-Hsiao, and Ting Hsiang Lin, 2017, "Modeling Data with a Truncated and Inflated Poisson Distribution." *Statistical Methods & Applications* 26(3): 383-401.

Wang, Hao, and Daniel F. Heitjan, 2008, "Modeling Heaping in Self-reported Cigarette Counts." *Statistics in Medicine* 27(19): 3789-3804.

Welsh, Alan H., Ross B. Cunningham, Christine F. Donnelly, and David B. Lindenmayer, 1996, "Modelling the Abundance of Rare Species: Statistical Models for Counts with Extra Zeros." *Ecological Modelling* 88(1): 297-308.

Zhou, Xiao-Hua, and Wanzhu Tu, 1999, "Comparison of Several Independent Population Means When Their Samples Contain Log-normal and Possibly Zero Observations." *Biometrics* 55(2): 645-651.