cjhr.gums.ac.ir
Caspian J Health Res. 2018;3(1):5-9 doi: 10.29252/cjhr.3.1.5

Caspian Journal of Health Research

Original article

Efficiency of Zero-Inflated Generalized Poisson Regression Model on Hospital Length of Stay Using Real Data and Simulation Study

Roghaye Farhadi Hassankiadeh[1], Anoshirvan Kazemnejad[1*], Mohammad Gholami Fesharaki[1], Siamak Kargar Jahromi[2]

[1]*Department of Biostatistics, Tarbiat Modares University, Tehran, Iran.*
[2]*Shariati Hospital, Medical Education Research Center, Tehran, Iran*

[*]*Corresponding author: Anoshirvan Kazemnejad*
Email: Kazem_an@modares.ac.ir

ABSTRACT

Background: An important feature of Poisson distribution is the equality of mean and variance. However, additional zeroes in the data may cause over-dispersion in most cases, in which zero-inflated models are recommended. In this study, we aimed to evaluate the efficacy of zero-inflated models to predict hospital length of stay (LOS) using real data and simulated study.
Methods: This study was conducted on patients admitted at Shariati hospital, Tehran, Iran. Zero inflated Poisson (ZIP), zero inflated negative binomials (ZINB) and zero inflated generalized Poisson (ZIGP) models were fitted on patient's length of stay. The fitted models were compared using the Akaike information criterion (AIC). The simulated data was generated using a model with the lowest AIC. Different models were then compared using the AIC. Data analysis was performed in R statistical software.
Results: The results of both real data and simulation study showed lower AIC for ZIGP model compared to ZIP and ZINB model.
Conclusion: Given the high dispersion and Zero Inflation in hospital LOS, the zero-inflated generalized Poisson regression model is the most suitable model to predict determinants of LOS.

Keywords: Computer Simulation, Hospital, Length of Stay, Poisson Distribution
Citation: Farhadi Hassankiadeh R, Kazemnejad A, Gholami Fesharaki M, Kargar Jahromi S. Efficiency of zero-inflated generalized poisson regression model on hospital length of stay using real data and simulation study. Caspian J Health Res. 2018;3(1):5-9. doi: 10.29252/cjhr.3.1.5

ARTICLEINFO

Received: March 30, 2017
Accepted: November 18, 2017
ePublished: March 04, 2018

Introduction

One of the main goals of applied statistics in medicine is the data modeling to explain and analyze the health and medical phenomena. Various distributions such as Poisson regression can lead to underestimation of the standard deviation or estimating significant predictive variables erroneously (2). Sometimes over-dispersion is due to the existence of many zeroes in the data (2). In this case, the problem is underestimation of probability of zero using the Poisson model. To overcome this situation, zero-inflated models are suggested (3). Zero-inflated models are a group of statistical

models based on probability distribution that allows for frequent zero-valued observations. Excess zero in count data are common in health and biomedical application. Length of stay (LOS) in hospital that is a useful indicator of resource consumption and hospital efficiency, may be characterized by over-abundance of zeroes and considered as zero-inflated data (4). The length of the patient's stay is considered as numerical data. If the mean is equal to variance, then Poisson regression can be used, and if the variance is greater than the mean, negative binomial regression or generalized Poisson regression should be used (5). In some numerical data, such as LOS, there may be accumulation in a subset of data that may result in over-dispersion. In this case, the use of negative binomial or generalized Poisson models cannot modify over-dispersion (6). In such situation, mixed models such as zero inflated models can be applied (5-8).

Because of limited application of zero inflated models in prediction of LOS, this study aimed to use advanced statistical methods to select the best model for accumulated zero data in patients' LOS using real data and simulation study

Methods

This cross-sectional study was performed on patients undergoing general surgery at Shariati Hospital in Tehran, Iran. Those patients who did not discharge from the hospital during the study period were excluded. A total of 220 subjects were randomly selected using simple random sampling method on patients' hospital record numbers. Demographic, clinical, and hospital characteristics affecting length of stay were selected using surgical expert panel. According to four experts' opinions, those factors which at least half of the experts considered them as effective factors were defined in data collection forms. Accordingly, out of 36 primary characteristics extracted from the previous researches, 32 features were selected. These features were classified based on the type of operation into thirteen groups (appendix, digestive system, abdominal and peritoneum, bone and muscular system, blood and lymphatic system, internal glands, urinary tract system, hemorrhoids, cardiovascular, female genital system, breast, lung, and skin). Other variables were the type of patient's insurance, history of smoking, types of disease (blood pressure, diabetes, blood lipids, kidney disease and other illnesses), age, gender, patient status at admission, the number of operations, and hemoglobin level.

The response variable was length of stay. There were many patients with one day hospitalization. We fitted the models in the present data minus 1 of the patients' length of stay. By doing this change, 54.5% had zero days of stay. Therefore, the advanced zero-inflated models can be used at zero point.

In zero-inflated models, the distribution of the response variable was composed of the combination of two zero and non-zero components. These models are named based on the distribution for the non-zero part of the model. For example, if the non-zero Poisson is considered, then it is called zero-inflated Poisson (ZIP) models, if the distribution of negative binomials was negative, then it is called zero-inflated negative binomial (ZINB) model and, if the Poisson distribution is considered as generalized, it would be called the zero-inflated generalized Poisson (ZIGP) model.

The formula of the zero-inflated model with D distribution (Poisson, binomial, generalized Poisson, etc.) is as follows (3).

$$P(X = X) = \begin{cases} \varphi + (1-\varphi) f_D(0) \\ (1-\varphi) f_D(x) \qquad x = 1,2,\ldots \end{cases}$$

Where, φ is the probability of zero, i.e. (0 <φ <1).

ZIP model is another type of Poisson distribution that can be used to fit count data. ZIP has different properties with Poisson distribution. In ZIP the probability of the presence of zero is higher than that in standard Poisson distribution. The mean and variance of ZIP include:

$$E(y_i \mid x_i, z_i) = \mu_i (1-\varphi_i)$$
$$\text{var}(y_i \mid x_i, z_i) = \mu_i (1-\varphi_i)(1+\mu_i \varphi_i)$$

The notable feature of this distribution is that the mean is higher than the variance. The zero-inflated Poisson regression model has common characteristics with Poisson distribution, and both models have a variance larger than the mean (9) .

ZINB model is a regression model that is fitted with a binomial distribution. The data are generated from a negative binomial model, and its mean and variance are as follows (10, 11):

$$E(y_i \mid x_i, z_i) = \mu_i (1-\varphi_i)$$
$$\text{var}(y_i \mid x_i, z_i) = \mu_i (1-\varphi_i)(1+\mu_i \varphi_i)$$

ZIGP model is defined as follows.

$$P(Y_i = y_i \mid x_i, z_i) =$$
$$\varphi_i + (1-\varphi_i) f(\mu_i, \alpha; 0) y_i = 0$$
$$(1-\varphi_i) f(\mu_i, \alpha; y_i) y_i > 0$$

Where, $y_i = 0,1,2,\ldots$ and $f(\mu_i, \alpha; y_i)$ are the density function of the generalized Poisson regression model (GP), and it is the probability of zero occurrence for the response variable. Therefore, ZIPG is a combination of Bernoulli distribution with $(1-\varphi)$ parameter and GP distribution with μ and α parameters. The mean and variance of the ZIGP model are obtained as follows(11,12):

$$E(y_i \mid x_i) = (1-\varphi_i)\mu_i$$
$$\text{var}(y_i \mid x_i) = (1-\varphi_i)(\mu_i^2 + \mu_i (1+\alpha\mu_i)^2) - (1-\varphi_i)^2 \mu_i^2$$
$$= E(y_i \mid x_i)\left((1+\alpha\mu_i)^2 + \varphi_i \mu_i\right)$$

The length of stay was then simulated based on the existing data parameters, and the best model was simulated in the actual data example. Then, the zero-inflated Poisson, zero-inflated negative binomial, and zero-inflated generalized Poisson regression models were fitted to these data. Simulation was conducted using the best model in the example of the patient's length of stay i.e. the generalized

zero-inflated Poisson regression model. In this case, one can assume that the data is a combination of distinct dual data in a production process. One of which produces only zero, and another one generates data from the Poisson or Binomial or generalized Poisson distribution. For example, for ith observation, the first phase (the zero component) is chosen with the probability of φ and the second phase (non-zero component) with the probability of 1-φ. The first phase only produces a zero count data, while the second phase i.e. the count data (non-zero component) produces a Poisson model or a negative binomial or generalized Poisson.

After simulation, the AIC and standard deviation were used to compare the fitted model with the observed values of the response variable and different zero inflated models.

$$AIC = -2\log L(\hat{\omega}) + 2p$$

Where ω is the maximum likelihood estimation of the model parameters, and p is the number of model parameters. The lower the size of this statistic, the more appropriate the model is. In this study, the R software (ZIGP Package) was used to analyze data.

Results

Of 220 patients, 87 (39.5%) were females and 133 (60.5%) were males. The mean age of patients was 44.35 (standard deviation=15.62). Most of the subjects (82.7%) lived in Tehran, and 83.6% had insurance. More than half of patients (54.5%) were hospitalized for one day and 18.1% for two days. The number of admission days minus 1 produced too many zeros. The variance of length of stay (12.16) was higher than the mean (1.76). Table 1 shows demographic and clinical characteristics of patients.

The estimation of significant coefficients of zero-inflated models on the patients' length of stay are shown in table 2. According to AIC, the best model was generalized zero-inflated Poisson regression model. According to the results, factors such as type of surgery can reduce the patients' length of stay. Having insurance, suffering from comorbidities, residence in Tehran, the number of tests, the number of operations, and age increased the patients' length of stay. Simulated data were generated using two different zero probability values. The results of each simulation of the length of stay data with 1000 repetitions (rep = 1000) are shown in table 3. According to the simulation results, the models were compared using the Standard Error (S.E) and AIC (smaller S.E and AIC, better fit). The highest values of the AIC belonged to data with the ZIP model, and ZIGP had the lowest AIC indicating that ZIGP model works better than the ZIP and ZINB, and ZIP model has the worst fit. In the second simulation with higher probability for zero values (0.6), the AIC decreased. Therefore the accuracy of the estimates rises by increasing the number of zeros.

Table 1. Demographic and clinical characteristics of the patients

Characteristics	Number	Percent
Type of operation		
Appendix	20	9.1
Digestion	76	34.5
Abdominal and Peritoneum and	32	14.5
Bone and muscular system	13	5.9
The blood and lymph system	10	4.5
Glands	31	14.1
Urinary tract	2	0.9
Hemorrhoids	15	6.8
Cardiovascular	16	7.3
Female reproductive system	2	.9
Breast	5	2.3
Lung	4	1.8
Skin	8	3.6
History of smoking	30	13.6
Comorbidities		
Hypertension	37	16.8
Diabetes	29	13.2
Heart disease	27	12.3
Hyperlipidemia	13	5.9
Kidney disorder	19	8.6
Other comorbidities	25	11.4
Anemia	159	72.3
Patient status at discharge		
Recovery	179	81.4
Personal satisfaction	34	15.5
Death	7	3.2
Number of tests		
0	5	2.3
1	11	5.0
2	28	12.7
3	41	18.6
4	44	20.0
5 and more	91	41.4
Number of operation		
1	153	69.5
2	38	17.3
3 and more	29	13.2
Gender		
Female	87	39.5
Male	133	60.5
Age groups (years)		
<14	2	.9
14-44	104	47.3
45-64	93	42.3
>65	21	9.5
Marital status		
Single	49	22.3
Married	171	77.7
Insurance		
No	36	16.4
Yes	184	83.6
Place of residence		
Non native	38	17.3
Native	182	82.7
Hospital stay length		
0	120	54.55
1	40	18.2
2	19	8.6
3 and more	41	18.6

Table 2. Estimation of the significant coefficients of ZIGP, ZINB and ZIP models

Variables	ZIP			ZINB			ZIGP		
	Estimate	SE	P-value	Estimate	SE	P-value	Estimate	SE	P-value
Appendix	-3.5	0.68	0.001	-3.19	0.54	0.01	-2.8	0.42	0.005
Abdominal & Peritoneum	-0.79	1.48	0.03	-0.80	1.36	0.02	-0.65	1.21	0.003
Hemorrhoids	-2.05	0.42	0.03	-1.50	0.35	0.03	-1.9	0.24	0.04
Lung	-.98	0.63	0.006	-1.56	0.59	0.04	-1.52	0.56	0.04
Skin	-2.14	0.65	0.03	-2.05	0.69	0.04	-2.17	0.64	0.03
Having insurance	0.44	0.87	0.02	0.36	0.85	0.03	0.52	0.79	0.04
Hypertension	0.49	0.18	0.001	0.52	0.25	0.02	0.46	0.18	0.03
Hyperlipidemia	0.53	0.033	0.009	0.47	0.32	0.04	0.56	0.21	0.04
Place of residence	-.047	0.34	0.001	-0.59	0.19	0.009	-0.39	0.22	0.01
Number of surgery	0.260	0.087	0.001	0.213	0.07	0.006	0.226	0.077	0.003
Number of tests	0.28	0.36	0.007	0.25	0.28	0.009	0.27	0.26	0.008
Age	0.18	0.43	0.03	0.32	0.34	0.04	0.28	0.22	0.04
Model AIC	921.132			809.2004			806.326		

Abbreviations: SE, Standard error; ZIP, Zero inflated Poisson model; ZINB, Zero-inflated Negative binomial model; ZIGP, Zero-inflated generalized Poisson model; AIC, Akaike information criterion

Table 3. The results of fitting the zero-inflated regression models using simulation

	Parameter	ZIP	ZINB	ZIGP
First simulation with $\varphi = .5$	φ	0.38	0.31	0.52
	S.E	0.344	0.246	0.161
	μ	3.41	6.9	1.14
	S.E	0.339	0.214	0.135
Comparing models	AIC	576.28	184.74	98.84
	Parameter	**ZIP**	**ZINB**	**ZIGP**
Second simulation with $\varphi = .6$	φ	.54	.142	.32
	S.E	0.299	0.143	0.091
	μ	2.57	7.38	1.8
	S.E	0.302	0.152	0.123
Comparing models	AIC	487.320	153.32	80.42

Abbreviations: ZIP, Zero inflated Poisson model; ZINB, Zero-inflated Negative binomial model; ZIGP, Zero-inflated generalized Poisson model; S.E, Standard Error; φ, Zero probability value in the data; μ, the mean of fitted model

Discussion

The findings of current study using real and simulated data revealed that zero inflated generalized Poisson model is the suitable model for predicting length of hospital stay. LOS as a linear outcome generally has been predicted using linear regression models (13-15). One of the important assumption for linear regression is the homogeneity of variance and the normal distribution of residuals. But similar to LOS, many of health and biomedical data do not have the conditions for using linear regression model (4). Hence, it is one of the topics of interest among researchers in various sciences such as medicine and economics to find new models and analyze discrete data such as generalized models (7, 8, 16-19). These models have an extra parameter which can estimate the extra zero values that are not well estimated with the help of the hypothesized model. Zero inflated models has also been used by some previous studies to model LOS. Wang et al. fitted a zero-inflated Poisson (ZIP) mixed model to identify health and patient related characteristics associated with LOS and to model variations in LOS within Diagnosis Related Groups. Their investigation revealed that age, number of diagnoses, and admission types are important risk factors affecting LOS (20). In a study by Sepehri et al. using the Vietnam National Health Survey and ZINB regression model, Vietnam's health insurance schemes has been identified as determinant of both hospital admission and LOS. In their results, compulsory health insurance and health insurance for the poor increased the expected LOS, while voluntary health insurance had minimal effect on the LOS (21). Rafeei and colleagues studied truncated numerical models in zero point with Poisson and negative binomial distributions as well as ZIP and ZINB regression models on the length of stay among mothers referred to health centers in Arak (7).

In the case where the response variable is numeric and non-negative, the proper model is Poisson or negative binomial regression (1, 20, 22). The assumptions of the conventional Poisson model constrain some count data (2, 23). The main condition for using the Poisson model is the equivalence of mean and variance of the response variable. If this condition is not present, the generalized Poisson distribution model will be appropriate (24-26). Discrete-count data sometimes show excessive dispersion and cannot be explained by a simple model such as binomial or Poisson. On the other hand, in some medical processes, there are too much zero, but the rest of the non-zero values follow the Poisson distribution. In other words, these data sometimes display a large number of zero which is more than what we expected; therefore, fitting general linear models such as binomial or Poisson model on a set of such data may not be acceptable.

The zero-inflated models have been proposed for a time when the data have large number of zeroes that might lead to over-dispersion of the data and violating the equivalence of mean and variance. In such situation, the zero-inflated models can well explain this over-dispersion.

In current study, we had an accumulation value of 1 that converted to zero-inflated through subtracting from 1. This transformation resulted in a variance of 12.16 and mean of 1.76 for LOS indicating over-dispersion of data. However, the use of negative binomial regression and generalized Poisson regression, has the ability to explain over-dispersion, but in cases where the number of zeros exceeds the limit or is so-called zero-inflated, these two strategies cannot explain and analyze the data. Therefore, one of the optimal methods to analyze the explanation of zero-inflated data zero is using regressions with zero-inflated distribution (2).

Conclusion

According to the results of the real data and simulation analysis, the zero inflated generalized Poisson regression model was the best model to fit length of stay. Therefore, the zero-inflated generalized Poisson regression model is recommended in case of over-dispersion data such as length of stay.

Acknowledgement

The authors acknowledge the cooperation of the personnel of Shariati hospital.

Ethical consideration

This study was reviewed and approved by Tarbiat Modarres University Research Ethics Board, Tehran, Iran.

Conflict of interests

The authors declare that they have no conflict of interests.

Funding

This study has been funded by Tarbiat Modarres University as student thesis.

References

1. Famoye F, Wulu JT, Singh KP. On the generalized Poisson regression model with an application to accident data. J Data Sci. 2004;2(3):287-295.
2. Hilbe JM. Negative binomial regression. Cambridge: Cambridge University Press; 2011.
3. Ntzoufras I, Katsis A, Karlis D. Bayesian assessment of the distribution of insurance claim counts using reversible jump MCMC. N Am Actuar J. 2005;9(3):90-108.
4. Feng C, Li L. Modeling Zero Inflation and Overdispersion in the Length of Hospital Stay for Patients with Ischaemic Heart Disease. In: Chen D, Chen J, Lu X, Yi G, Yu H, eds. Advanced statistical methods in data science. Singapore: Springer; 2016.
5. Böhning D, Dietz E, Schlattmann P, Mendonça L, Kirchner U. The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. J R Stat Soc Ser A Stat Soc. 2000;163(1):121-122.
6. Hall DB. Zero-inflated Poisson and binomial regression with random effects: a case study. Biometrics. 2000;56(4):1030-1039.
7. Rafiee M, Ayatollahi MT, Behboodian J. Zero-inflated negative binomial modeling, efficiency for analysis of length of maternity hospitalization [in Persian]. Yafteh. 2005;6(4):47-58.
8. Ridout M, Hinde J, DeméAtrio CG. A Score Test for Testing a Zero-Inflated Poisson Regression Model Against Zero-Inflated Negative Binomial Alternatives. Biometrics. 2001;57(1):219-223.
9. McCullagh P, Nelder JA. Generalized Linear Models. 2nd ed. London: Chapman & Hall; 1983.
10. Allison PD. Comparing logit and probit coefficients across groups. Sociol Methods Res. 1999;28(2):186-208.
11. Czado C, Erhardt V, Min A, Wagner S. Zero-inflated generalized Poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcing rates. Stat Model. 2007;7(2):125-153.
12. Jiménez RE, Lam RM, Marot M, Delgado A. Observed-predicted length of stay for an acute psychiatric department, as an indicator of inpatient care inefficiencies. Retrospective case-series study. BMC Health Serv Res. 2004;4(1):4.
13. Jaffee EG, Arora VM, Matthiesen MI, Meltzer DO, Press VG. Health Literacy and Hospital Length of Stay: An Inpatient Cohort Study. J Hosp Med. 2017;12(12):969-973.
14. Passias PG, Jalai CM, Worley N, et al. Predictors of Hospital Length of Stay and 30- Day Readmission in Cervical Spondylotic Myelopathy Patients: An Analysis of 3057 Patients Using the ACS-NSQIP Database. World Neurosurg. 2018; 110: e450-e458.
15. Valent F, Tonutti L, Grimaldi F. Does diabetes mellitus comorbidity affect in-hospital mortality and length of stay? Analysis of administrative data in an Italian Academic Hospital. Acta Diabetol. 2017; 54(12):1081-1090.
16. 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.
17. Greene WH. Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. Available at: https://ssrn.com/abstract= 1293115. Updated November 03,2008. Accessed January 12, 2018.
18. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics. 1992;34(1): 1-14.
19. Ridout M, Demétrio CG, Hinde J. Models for count data with many zeros. Presented at: Proceedings of the XIXth international biometric conference. Cape Town, South Africa, December 14-18, 1998.
20. Wang K, Yau KK, Lee AH. A zero-inflated Poisson mixed model to analyze diagnosis related groups with majority of same-day hospital stays. Comput Methods Programs Biomed. 2002;68(3):195-203.
21. Sepehri A, Simpson W, Sarma S. The influence of health insurance on hospital admission and length of stay--the case of Vietnam. Soc Sci Med. 2006;63(7):1757-1770.
22. Cameron AC, Trivedi PK. Regression analysis of count data. Vol 53. Cambridge: Cambridge university press; 2013.
23. Agresti A. Categorical data analysis. 3rd ed. Hoboken: Wiley; 2014.
24. Karlis D, Xekalaki E. Mixed poisson distributions. Int Stat Rev. 2005;73(1):35-58.
25. Ng S, Yau K, Lee A. Modelling inpatient length of stay by a hierarchical mixture regression via the EM algorithm. Math Comput Model. 2003;37(3-4):365-375.
26. Wulu J, Singh K, Famoye F, Thomas T, McGwin G. Regression analysis of count data. Jour Ind Soc Ag Statistics. 2002;55(2):220-231.