Zero-Inflated Poisson Models

Elena Volpi

December 7, 2022

1 Introduction and Motivation

In the study of general regression models, we often run into the problem of excessive zeroes with count data. This problem is called over-dispersion and usually does not have an obvious solution. Research has shown that zero-inflated Poisson models may not adequately address over-dispersion and zero-inflated negative binomial regression parameter estimation may fail to converge. We will examine the proposed model by Felix Famoye and Karen P. Singh in *Zero-Inflated Generalized Poisson Regression Model with an Application to Domestic Violence Data* as a possible alternative.

2 Models

2.1 Generalized Poisson Model

Let y_i be the number of events that occurred. Then the generalized Poisson model (GPR) is:

$$f(\mu_i, \alpha; y_i) = (\frac{\mu_i}{1+\alpha\mu_i})^{y_i}\frac{(1+\alpha y_i)^{y_i-1}}{y_i!}\exp[\frac{-\mu_i(1+\alpha y_i)}{1+\alpha\mu_i}] \text{ , where} \tag{1}$$

$$y_i = 0, 1, 2, ...;\ i = 1, 2, ..., n;\ \mu_i = \exp(\sum x_{ij}\beta_j),\ \ x_i \text{ is the i-th row of the covariate matrix } \mathbf{X}$$

This gives link functions of Our dispersion parameter is α, when $\alpha = 0$ we have equi-dispersion and the model reduces to the Poisson regression model.

2.2 Zero-Inflated Generalized Poisson Model

Let out zero-inflated generalized Poisson model (ZIGP) be defined as:

$$P(Y = y_i|x_i, z_i) = \varphi_i + (1-\varphi_i)f(\mu_i, \alpha; 0), \qquad , y_i = 0 \tag{2a}$$

$$= (1-\varphi_i)f(\mu_i, \alpha; 0), \qquad , y_i > 0 \tag{2b}$$

,where $f(\mu_i, \alpha; y_i), y_i = 0, 1, 2, ...$ is the GPR model (1) and $0 < \varphi < 1$, $\mu_i = \mu_i(x_i)$, $\varphi_i = \varphi_i(z_i)$, where z_i is the i-th row of the covariate matrix $\mathbf{Z}$. Assuming y_i is independent and following a zero-inflated generalized Poisson distribution, zeroes occur in two states, shown in equation 2a. With probability φ_i, we have structural zeroes and with probability $(1-\varphi_i)$ we have sampling zeroes. The sampling zeroes lead to a generalized Poisson distribution with parameters α and φ_i.

The expectation and variance formulas for the ZIGP are as follows:

$$E(y_i|x_I) = (1-\varphi_i)\mu_i(x_i) \tag{3}$$

$$
\begin{aligned}
Var(y_i|x_I) &= (1-\varphi_i)[\mu_i^2 + \mu_i(1+\alpha\mu_i)^2] - (1-\varphi_i)^2\mu_i^2 \\
&= E(y|x_i)[(1+\alpha\mu_i)^2 + \varphi_i\mu_i]
\end{aligned} \tag{4}
$$

We have over-dispersion when $\varphi_i > 0$ and the model reduces to the GPR when $\varphi_i = 0$ Our links functions are,

$$
\log(\mu_i) = \sum_{j=1}^{k} x_{ij}B_j \text{ for } \mu_i = \mu_i(x_i) \tag{5a}
$$

$$
\text{logit}(\varphi_i) = \log(\varphi_i[1-\varphi_i])^{-1} \tag{5b}
$$

Furthermore, if the same covariates effect φ_i and μ_i, we may write our link functions as:

$$
\log(\mu_i) = \sum_{j=1}^{k} x_{ij}\beta_j \tag{6a}
$$

$$
\text{logit}(\varphi_i) = \log(\frac{\varphi_i}{1-\varphi_i}) = -\tau\sum_{j=1}^{k} x_{ij}\beta_j \tag{6b}
$$

, this is the ZIGP(τ) model. When $\alpha = 0$, ZIGP(τ) reduces to Zero-Inflated Poisson ZIP(τ), from Lambert (1992). For both models, when $\tau > 0$, zeros become more likely. Parameter estimation may be done through maximum likelihood estimation, in particular, the author's utilized Newton-Raphson

3 Motivating example

In part of a plan to reduce domestic violence in Portland, Oregon the Family Violence Intervention Steering Committee of Multnomah County and Portland Police Bureau conducted a study utilizing records on batterers and victim surveys from 1996-1997. The data used in this project is the second wave of interviews conducted six months after the recorded police case for each survivor.

3.1 Variables

Survey responses from the survivor were recorded and assailant information was drawn from police records. Survey questions related to violence incidents in the past six months were aggregated into a single variable, Violence, which is the number of violent behaviors toward the victim. We have a variable response for both the survivor and assailant for variables Education, Income, Employment, Family, Club and Drug. Education had response 1-3 (1: some high school or less, 2: High school diploma or GED, 3: Some college of more) and Income was 1-4, each representing a bracket of income where (\$0 - \$5k, \$5k - \$10k, \$20k-\$30k, > \$30k). Our indicator variables are full time employment, interact with family, belong to a club, and have a drug problem. After removing missing responses, we have 214 cases.

Table 1: Descriptive statistics for the variables

Variable	Description	Mean ± SD	Proportion of 1's
Edu_v	Education level, victim	2.2897± 0.7507	
Edu_b	Education level, batterer	2.0654± 0.7785	
Emp_v	Full time employment, victim		0.5047
Emp_b	Full time employment, batterer		0.6589
Inc_v	Income level, victim	2.5654±1.3083	
Inc_b	Income level, batterer	3.0701±1.4727	
Fam_v	Interact with family, victim		0.8224
Fam_b	Interact with family, batterer		0.7196
Club_v	Belong to a club, victim		0.2710
Club_b	Belong to a club, batterer		0.1916
Drug_v	Have drug problem, victim		0.1355
Drug_b	Have drug problem, batterer		0.6215
Violence	Number of domestic violence	4.2056±10.6014	

In the regression setting, our response y_i is the number of violent incidents that occurred. The observed proportion of zeros is 66.4% in the domestic violence data.

4 Model Comparison

The iterative approach for the zero-inflated negative binomial model parameter estimation failed to converge on several cases. The General Poisson Model can be tested on the data using a score test. The process of the score test must be skipped for brevity. Under the null hypothesis, the score statistic has an asymptotic chi-sqared distribution and was found to be $20.02 \sim X_1^2$. The score statistic is significant at the 5% significant levels leading us to conclude that the domestic violence data has too many zeros for the general Poisson model to be adequate. Finally, we'll compare ZIGP(τ) model to the ZIP(τ) model on the domestic violence data using a goodness of fit test. The estimated proportion of zeros from ZIP and ZIGP are 63.7% and 65.7%. Let us consider $H_0 : \alpha = 0$ vs $H_1 : \alpha \neq 0$. The table below shows the results of the goodness of fit test using the asymptotic Wald statistic:

Table 3: Estimates from ZIP regression and ZIGP regression models

	ZIP		ZIGP	
Variable	Estimate ± SE	t-value	Estimate ± SE	t-value
Intercept	3.4206 ± 0.1729	19.78**	5.4332 ±1.2620	4.31**
Edu_v	-0.3569 ± 0.0550	-6.49**	-1.5005 ± 0.4967	-3.02**
Edu_b	0.0370 ± 0.0527	0.70	0.5907 ± 0.3035	1.95
Emp_v	0.1252 ± 0.0897	1.40	0.3419 ± 0.5027	0.68
Emp_b	0.0211 ± 0.1051	0.20	1.2458 ± 0.7711	1.62
Inc_v	-0.0878 ± 0.0362	-2.43*	-0.4814 ± 0.2154	-2.24*
Inc_b	-0.2012 ± 0.0384	-5.25**	-0.4183 ± 0.2466	-1.70
Fam_v	0.1245 ± 0.0999	1.25	0.1804 ± 0.4629	0.39
Fam_b	-0.1645 ± 0.0696	-2.36*	-0.6656 ± 0.4951	-1.34
Club_v	0.7804 ± 0.1050	7.43**	1.7158 ± 0.7047	2.43*
Club_b	-0.8548 ± 0.1222	-7.00**	-1.9866 ± 0.7128	-2.79**
Drug_v	-0.7577 ± 0.1275	-5.94**	-1.0645 ± 0.5377	-1.98*
Drug_b	0.6305 ± 0.0929	6.79**	1.5428 ± 0.4019	3.84**
τ	-0.2456 ± 0.0619	-3.97**	-0.1242 ± 0.0570	-2.18*
α			0.3050 ± 0.0556	5.49**
Log-likelihood	-641.09		-365.84	

* indicates significant at 0.05 level; ** indicates significant at 0.01 level; SE = standard error

We see strong evidence that α is significantly different from zero which leads us to conclude that the preferable model for this data is the ZIGP model. The ZIGP has nearly a twice as large log-likelihood.

5 Conclusion

The zero-inflated negative binomial model may be a good competitor but , as in this case, parameter estimation may fail to converge. The zero inflated Poisson model and generalized Poisson did not adequately address over-dispersion. The domestic violence example shows the usefulness of a zero-inflated generalized Poisson model.

References

[1] Famoye, Felix Singh, Karan. (2006). Zero-inflated generalized Poisson regression model with an application to domestic violence data. Journal of Data Science. 4. 117-130.

[2] Jolin, Annette, Fountain, Robert, Feyerherm, William H., and Friedman, Sharon. Portland [Oregon] Domestic Violence Experiment, 1996-1997. Inter-university Consortium for Political and Social Research [distributor], 2006-07-24. https://doi.org/10.3886/ICPSR03353.v2

[3] Lambert, Diane. Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. Technometrics, vol. 34, no. 1, 1992, pp. 1–14. JSTOR, https://doi.org/10.2307/1269547. Accessed 8 Dec. 2022.