Diving into the Deep End of Machine Learning: Using Keras for Predicting Education Outcomes with Tabular Data

Elena Volpi

March 14, 2023

Statistic Master’s Project

Background and Motivation

Paper: ”Zero-Inflated Generalized Poisson Regression Model with an Application to Domestic Violence Data”
Authors: Felix Famoye and Karen P. Singh

Background and Motivation

Motivation

1. Authors’ previous research exposed not all zero-inflated Poisson models were adequately addressing overdispersion of count data.
2. Also, Zero-inflated negative binomial regression was also inadequate. The iteration for estimating parameters failed to converge.

About the data

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 victim.

Variables

Survey Responses were recorded for batterer and victim. Missing responses removed.

- Violence: The number of violent behaviors toward the victim
- Education:
  1. - Some high school or less
  2. - High School diploma or GED
  3. - Some college or more
- Income :
  1. - $0 - $5k
  2. - $5k - $10k
  3. - $20k - $30k
  4. > $30k
- Binary Response
  - Full time employment
  - Interact with family
  - Belong to a club
  - Have a drug problem

About the Data

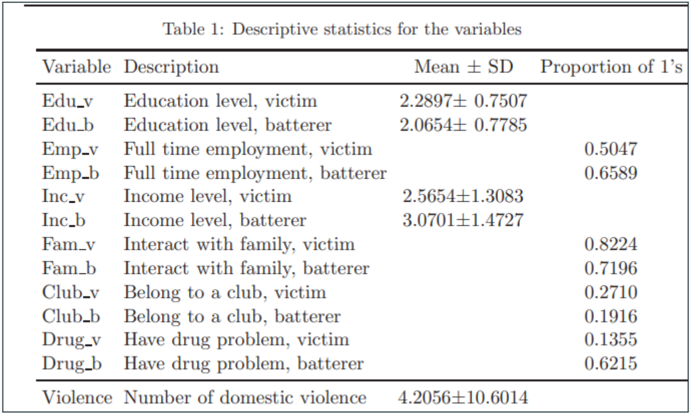

Table 1: Descriptive statistics for the variables

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	

There are 214 cases after removing cases with missing information.

Generalized Poisson Model

Let y_i be the number of violent behaviors that occurred towards the victim. Then the generalized poisson model (GPR) is:

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

, where

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

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

Zero-Inflated Generalized Poisson Model

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

$$\begin{aligned} P(Y = y_i|x_i, z_i) &= \varphi_i + (1 - \varphi_i)f(\mu_i, \alpha; 0), && , y_i = 0 \\ &= (1 - \varphi_i)f(\mu_i, \alpha; 0), && , y_i > 0 \end{aligned}$$

, where $f(\mu_i, \alpha; y_i), y_i = 0, 1, 2, ...$ is the GPR model and $0 < \varphi < 1$

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

We have overdispersion when $\varphi_i > 0$ and the model reduces to the GPR when $\varphi_i = 0$

Log-Link function: $log(\mu_i) = \sum_{j=1}^{k} x_{ij} B_j$ for $\mu_i = \mu_i(x_i)$
Logit link: $logit(\varphi_i) = log(\varphi_i[1 - \varphi_i])^{-1}$

If the same covariates effect φ_i and μ_i, we have:

$$ZIGP(\tau) = log(\mu_i) = \sum_{j=1}^{k} x_{ij}\beta_j, \;\; logit(\varphi_i) = log(\frac{\varphi_i}{1 - \varphi_i}) = -\tau \sum_{j=1}^{k} x_{ij}\beta_j$$

, when τ > 0 excess zeros are more likely.

When $\alpha = 0$, ZIGP(τ) reduces to ZIP(τ), Zero-Inflated Poisson from Lambert et al.

*"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 1 Dec. 2022.

Model Comparison

We can compare the ZIGP(τ) to the ZIP(τ).

- Maximum likelihood estimates: estimate using Newton-Raphson
- Score test

  Score statistic 20.02 $\sim X_1^2$, significant at 5% level thus GPR is not adequate.
- Observed proportion of zeros is 66.4% in domestic violence data
- Estimated proportion of zeros from ZIP and ZIGP are 63.7% and 65.7%

Results

Goodness of fit test: Test ZIGP adequacy over ZIP model using $H_0 : \alpha = 0$ vs $H_1 : \alpha \neq 0$ (Wald)

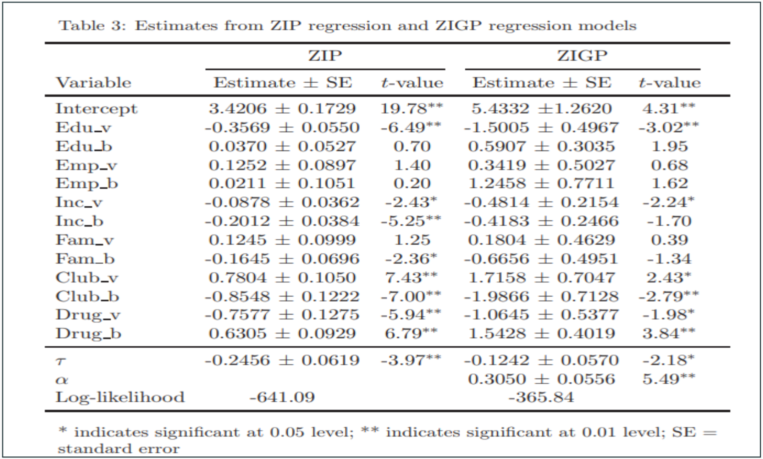

Table 3: Estimates from ZIP regression and ZIGP regression models

	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

Results

- Goodness of fit test conclusion: α significantly different from zero.

  ZIGP model fits better than ZIP

- ZIP : 6 independent variables significant at 1% level
- ZIGP: 3 independent variables significant at 1% level

Conclusion

- ZIGP regression successfully fitted to all datasets tested
- In a few cases, estimation of parameters of ZINB regression did not converge

”Even though the ZIGP regression model is a good competitor of ZINB regression model, we do not know under what conditions, if any, which one will be better. . . The application of the ZIGP regression model to the domestic violence data illustrates the usefulness of the model.” (pg 128)

Thank you.