پانزدهمین سمینار احتمال
و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان

15th Seminar On Probability and Stochastic Processes
August 30-31, 2025
University of Kurdistan

انجمن آمار ایران

Estimation of count parameters in zero-inflated Poisson regression model based on Stein-Liu estimators

Zahra Zandi*and Hossein Bevrani

Department of Statistics, University of Kurdistan, Sanandaj, Iran

Abstract

The ZIP regression is applied for modelling count data with extra zeros. In this paper, we have improved the count parameter estimation in this model with correlated count predictor variables under the assumption of sparseness of the model using Stein-Liu estimators. We have conducted a Monte Carlo simulation study for various combinations of sample sizes, inactive count predictor variables, and correlation level between the count predictor variables in order to compare the performance of the suggested estimator with the unrestricted Liu estimator in terms of simulated relative efficiency criteria. Our results showed that the performance of the Stein-Liu estimators is superior to the unrestricted Liu estimator in any situations.

Keywords: Monte Carlo simulation, Multicollinearity problem, Stein-Liu estimators, zero-inflated Poisson model
Mathematics Subject Classification (2020): 62F10, 62F12, 62J07.

1 Introduction

The Poisson regression is widely used to model equivalent count data, i.e. the average and variance of data are equal. However, in many real applications, the count data contains an excess number of zeros which is known as zero-inflated count data. Zero-inflated Poisson (ZIP) regression model introduced by Lambert (1992) is a two-part model to analyze this data. The first part generates only structural zeros using a degenerate distribution at zero. This part is called zero-component. The other part is called count component which generates true zeros and positive count data from an ordinary Poisson distribution. The interested readers are refereed to Hilbe (2011) for more details about ZIP model.

The multicollinearity problem that occurs in analysis count models includes correlated predictor variables. This problem leads to high variance of the maximum likelihood estimator (MLE) of the regression parameters. Thus, the interpretations based on this estimator can not be correct. The ridge regression estimator (Hoerl and Kennard (1970)) is a solution to overcome this problem. Also, the Liu estimator introduced by Liu (1993) is another method

*Corresponding author, zahrazandi163@yahoo.com

dealing with multicollinearity problem which has more advantages over the ridge estimator when there is a strong correlation between predictor variables.
When there is uncertain prior information (UPI) about some predictor variables which indicate that they are not significant and should be eliminated from regression model in order to improve parameters estimation, the Stein and positive Stein estimators introduced by Stein (1956) perform better than the MLE. The UPI appears on the parameter vector as a linear restriction (Ahmed (2014)).
In this paper, we consider a ZIP model in the presence of multicollinearity among the count predictors under a linear restriction on the count parameter vector $\boldsymbol{\beta}$ as

$$\mathbf{R}\,\boldsymbol{\beta} = \mathbf{r} \tag{1}$$

where, $\mathbf{R}_{p_2\times p}$ is a known matrix, $\boldsymbol{\beta}_{p\times 1} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ is the count parameters vector ($\boldsymbol{\beta}_{1\,(p_1\times 1)}'$ contains significant or active parameters and $\boldsymbol{\beta}_{2\,(p_2\times 1)}'$ includes non-significant or inactive parameters) and $\mathbf{r}_{p_2\times 1}$ is a vector with constant elements. Our mean goal is to improve count parameters in the ZIP model using Stein-Liu (SL) and positive Stein-Liu (PSL) estimators. Recently these estimators are proposed in the multicollinear Poisson model by Zandi and Bevrani (2024).
This paper is organized as follows. The ZIP model, unrestricted Liu (UL), restricted Liu (RL) and Stein-Liu estimators are introduced in Section 2. The SL estimators are compared to the UL estimator using a Monte Carlo simulation study in Section 3, and conclusions are presented in the final section.

2 Model specifications and estimators

In the current section, we introduce the ZIP model, the SL, and PSL estimators to improve the count parameters estimation in this model.

2.1 Zero-inflated Poisson model

Suppose Y be dependent variable in the ZIP model. The probability mass function of Y for a random sample of size n is as follows:

$$f_{Y_i}(y_i \mid \tau_i, \mu_i) = \begin{cases} \tau_i + (1-\tau_i)\, e^{-\mu_i} & y_i = 0 \\ (1-\tau_i)\left(\frac{\mu_i^{y_i}\, e^{-\mu_i}}{y_i!}\right) & y_i = 1, 2, 3, \ldots \end{cases}$$

where $\log(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$ is the link function of the Poisson regression model in the count component of ZIP model where, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})'$ is the ith row of design matrix $\mathbf{X}_{n\times p}$ with p count predictor variables and $logit(\tau_i) = \boldsymbol{z}_i'\boldsymbol{\gamma}$ is the link function related to the zero component, where $\boldsymbol{z}_i = (z_{i1}, z_{i2}, ..., z_{iq})'$ is the ith row of deign matrix $\mathbf{Z}$ and $\boldsymbol{\gamma}$ is a $q \times 1$ vector of parameters related to the zero component. The log-likelihood function of the ZIP model can be written as follows:

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \Big\{ I_i \ln\left(e^{\mathbf{z}_i'\boldsymbol{\gamma}} + e^{-\exp(\mathbf{x}_i'\boldsymbol{\beta})}\right) + (1-I_i)\Big(y_i\,\mathbf{x}_i'\boldsymbol{\beta} - e^{\mathbf{x}_i'\boldsymbol{\beta}} - \ln(y_i!)\Big) - \ln\left(1 + e^{\mathbf{z}_i'\boldsymbol{\gamma}}\right)\Big\}, \tag{2}$$

where I_i is an indicator function defined as:

$$I_i = \begin{cases} 1 & y_i = 0 \\ 0 & y_i = 1, 2, 3, \ldots. \end{cases}$$

We denote the MLE of the parameter vector $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ by $(\hat{\boldsymbol{\beta}}^{MLE}, \hat{\boldsymbol{\gamma}}^{MLE})$ that can be obtained by maximizing Equation (2) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ using numerical methods.

2.2 Unrestricted Liu and restricted Liu estimators

When there is near linear dependence among the count predictor variables in the ZIP model, the interpretations based on $\hat{\boldsymbol{\beta}}^{MLE}$ are not true. One solution is to use Liu estimator (or unrestricted Liu estimator) as follows (Liu (1993)):

$$\hat{\boldsymbol{\beta}}^{UL} = \left(\mathbf{X}'\hat{\boldsymbol{\mathcal{W}}}\mathbf{X} + \mathbf{I}_p\right)^{-1}\left(\mathbf{X}'\hat{\boldsymbol{\mathcal{W}}}\mathbf{X} + d\,\mathbf{I}_p\right)\hat{\boldsymbol{\beta}}^{MLE}, \qquad 0 \leq d \leq 1 \tag{3}$$

where $\hat{\boldsymbol{\mathcal{W}}}$ s a diagonal matrix such that the ith diagonal element is equal to $\hat{\mu}_i = exp(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}^{MLE})$, $\mathbf{I}_p$ is an identity matrix of order p and d is the shrinkage parameter. The MSE of the UL estimator is defined as follows (Gelman and Golam Kibria (2020)):

$$MSE(\hat{\boldsymbol{\beta}}^{UL}) = \sum_{j=1}^{p}\frac{(\lambda_j + d)^2}{\lambda_j(\lambda_j+1)^2} + (d-1)^2\sum_{j=1}^{p}\frac{\alpha_j^2}{(\lambda_j+1)^2} \tag{4}$$

where λ_j is the jth eigenvalue of matrix $\mathbf{C} = \mathbf{X}'\hat{\boldsymbol{\mathcal{W}}}\mathbf{X}$ and α_j is the jth element of $\mathbf{H}'\boldsymbol{\beta}$, where $\mathbf{H}$ is a $p \times p$ matrix whose its jth column is the corresponding eigenvector of λ_j where, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. The estimation of shrinkage parameter d is as follows (Månsson (2013)):

$$\hat{d} = max\Big(0, median\Big\{\frac{\hat{\alpha}_j^2 - 1}{\lambda_j^{-1} + \hat{\alpha}_j^2}\Big\}\Big),$$

where $\hat{\alpha} = (\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_p)' = \mathbf{H}'\hat{\boldsymbol{\beta}}^{MLE}$.
Under the linear restriction (1) and in the presence of multicollinearity, the RL estimator of $\boldsymbol{\beta}$ denoted by $\hat{\boldsymbol{\beta}}^{RL}$ is defined as follows (Gelman and Golam Kibria (2020)):

$$\hat{\boldsymbol{\beta}}^{RL} = \mathbf{A}\hat{\boldsymbol{\beta}}^{MLE} - \mathbf{A}\mathbf{C}^{-1}\mathbf{R}'\left(\mathbf{R}\mathbf{C}^{-1}\mathbf{R}'\right)^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}^{MLE} - \mathbf{r}), \tag{5}$$

where $\mathbf{A} = \left(\mathbf{X}'\hat{\boldsymbol{\mathcal{W}}}\mathbf{X} + \mathbf{I}_p\right)^{-1}\left(\mathbf{X}'\hat{\boldsymbol{\mathcal{W}}}\mathbf{X} + d\,\mathbf{I}_p\right)$.

2.3 Stein-Liu estimators

Based on the idea of Zandi and Bevrani (2024), the Stein-Liu estimator of $\boldsymbol{\beta}$ denoted by $\hat{\boldsymbol{\beta}}^{SL}$ is defined by combining the $\hat{\boldsymbol{\beta}}^{UL}$ and $\hat{\boldsymbol{\beta}}^{RL}$ as follows:

$$\hat{\boldsymbol{\beta}}^{SL} = \hat{\boldsymbol{\beta}}^{RL} + \Big(1 - \frac{p_2 - 2}{T_n}\Big)\Big(\hat{\boldsymbol{\beta}}^{UL} - \hat{\boldsymbol{\beta}}^{RL}\Big), \qquad p_2 = 3, 4, \ldots,$$

where T_n is the test statistic for testing $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ versus $H_1 : \mathbf{R}\boldsymbol{\beta} \neq \mathbf{r}$ which is defined as follows:

$$T_n = 2\Big\{\mathcal{L}(\hat{\boldsymbol{\beta}}^{UL}) - \mathcal{L}(\hat{\boldsymbol{\beta}}^{RL})\Big\},$$

where $\mathcal{L}(\hat{\boldsymbol{\beta}}^{UL})$ and $\mathcal{L}(\hat{\boldsymbol{\beta}}^{RL})$ are, respectively, the log-likelihood function (2) values for the unrestricted Liu and restricted Liu estimators. The test statistic T_n under H_0 has $\chi^2(p_2)$ when $n \to \infty$.
The Positive Stein-Liu estimator of $\boldsymbol{\beta}$ denoted by $\hat{\boldsymbol{\beta}}^{PSL}$ is defined as an adjusted version of the SL estimator:

$$\hat{\boldsymbol{\beta}}^{PSL} = \hat{\boldsymbol{\beta}}^{RL} + max\Big\{0, 1 - \frac{p_2 - 2}{T_n}\Big\}\Big(\hat{\boldsymbol{\beta}}_{UL} - \hat{\boldsymbol{\beta}}^{RL}\Big), \qquad p_2 = 3, 4, \ldots.$$

3 Simulation experiment

We compare the performance of the Stein-Liu estimators with respect to the UL in ZIP model in the presence of multicollinearity via a Monte Carlo simulation using the simulated relative efficiency (SRE) criteria which is defined as:

$$SRE(\hat{\boldsymbol{\beta}}^{UL}, \hat{\boldsymbol{\beta}}^{\diamond}) = \frac{SMSE(\hat{\boldsymbol{\beta}}^{UL})}{SMSE(\hat{\boldsymbol{\beta}}^{\diamond})},$$

where $\hat{\boldsymbol{\beta}}^{\diamond}$ is any one of the $\hat{\boldsymbol{\beta}}^{SL}$ and $\hat{\boldsymbol{\beta}}^{PSL}$, and $SMSE(\hat{\boldsymbol{\beta}}^{\diamond})$ is the simulated mean squared error of $\hat{\boldsymbol{\beta}}^{\diamond}$ which is defined as:

$$SMSE(\hat{\boldsymbol{\beta}}^{\diamond}) = \frac{1}{2000}\sum_{t=1}^{2000}(\hat{\boldsymbol{\beta}}^{\diamond} - \boldsymbol{\beta})_t'(\hat{\boldsymbol{\beta}}^{\diamond} - \boldsymbol{\beta})_t.$$

For the value of SRE is greater than one, $\hat{\boldsymbol{\beta}}^{\diamond}$ dominates the UL estimator. The correlated count predictor variables in the ZIP model were generated as:

$$x_{ij} = (1-\rho^2)^{1/2}\, u_{ij} + \rho\, u_{ip}, \quad i = 1, 2, ..., n, \quad j = 1, 2, ..., p,$$

here x_{ij}s are elements of design matrix $\mathbf{X}$, ρ represents the correlation level between the count predictor variables and u_{ij}s are generated independently from standard normal distribution. Hence, the response variable Y_i in the ZIP model is generated with $\mu_i = e^{\mathbf{x}_i'\boldsymbol{\beta}}$ and $\tau_i = 0.5$. The simulation is conducted in R statistical software under different sample sizes, $n = 35, 50$, various correlation level between the predictor variables, $\rho = 0.90, 0.93, 0.95$, and the inactive predictor variables, $p_2 = 3, 5, 7$. We also set the number of active predictors $p_1 = 3$, thus $p = p_1 + p_2 = 6, 8, 10$. The simulation is replicated 2000 times for each case. We consider the linear restriction as $\mathbf{R}\boldsymbol{\beta} = \mathbf{0}$, where $\mathbf{R} = [\mathbf{0}_{p_1\times p_2}, \mathbf{I}_{p_2\times p_2}]$ such that $\mathbf{0}_{p_1\times p_2}$ is a matrix with zero elements and $\mathbf{I}_{p_2\times p_2}$ is an identity matrix. Therefore, the count parameter is partitioned as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$, where $\boldsymbol{\beta}_1 = (0.09, 0.27, 0.34)'$ contains the active parameters and $\boldsymbol{\beta}_2 = (\mathbf{0}_{p_2})'$ includes the inactive parameters. To compare the performance of the Stein-Liu estimators to the UL estimator in different parts of the parameter space, we define a distance between the proposed sub-model and the simulated model as $\Delta = \|\boldsymbol{\beta} - \boldsymbol{\beta}^0\|^2$, where $\boldsymbol{\beta}^0 = (\boldsymbol{\beta}_1', \mathbf{0}_{p_2}')$ is the true parameter and $\|.\|$ is the Euclidean norm. We set $\Delta = 0, 0.4, 0.6, 0.8, 1, 2, 3, 4$ in the simulations.

The simulation results reported in Tables 1-3 confirm that the SREs of the RL and Stein-Liu estimators increase by increasing the number of inactive count predictors p_2 and also as the correlation level ρ increases. All these estimators have better performance than the UL estimator when the sample size n decreases. The restricted Liu estimator has the best performance at $\Delta = 0$. When Δ moves away from zero, the SRE of this estimator sharply decreases. The positive Stei- Liu estimator outperforms better than the Stein-Liu estimator at $\Delta = 0$ and near it in any situation. An important result of the proposed estimators is that the Stein-Liu estimators uniformly outperform the unrestricted Liu estimator.

Discussion

In this article, we proposed Stein-Liu and positive Stein-Liu estimators to estimate count parameters in Zero-inflated Poisson model in the presence of multicollinearity under a linear restriction on the count parameters. The performance of the suggested estimators was investigated through a Monte Carlo simulation study. The simulation results revealed that the performances of the Stein-Liu estimators were better than the unrestricted Liu estimator in various parts of the parameter space.

Acknowledgment

The authors are thankful to the reviewers for the insightful comments and suggestions that have resulted in a much improved version of this paper.

References

Ahmed S. E. (2014), *Penalty, shrinkage and pretest strategies – Variable selection and estimation*. Heidelberg, Springer.

Hilbe J. M. (2011), *Negative binomial regression*, Cambridge UK: Cambridge University Press.

Hoerl A. E. and Kennard R. W. (1970), Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics* **12 (1)**, 55–67

Gelman K. and Golam Kibria B. M. (2020), Estimating the unrestricted and restricted Liu estimators for the Poisson regression model: method and application. *Computational Economics* **58**, 311–326.

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

Liu K. (1993), A new class of biased estimate in linear regression. *Communications in Statistics-Theory and Methods* **22**, 393–402.

Månsson K. (2013), Developing a Liu estimator for the negative binomial regression model. *Journal of Statistical Computation and Simulation* **83** 1773–1780.

Stein C. (1956), The admissibility of Hotelling's T^2-test. *Mathematical Statistics* **27**, 616–623.

Zandi Z. and Bevrani H. (2024), Improved parameters estimation in the multicollinear Poisson regression model based on Stein-Liu estimators. *Journal of Statistical Modelling: Theory and Applications* **5 (2)**, 77–95.

Table 1: The *SREs* of the proposed estimators with respect to $\hat{\boldsymbol{\beta}}^{UL}$ for $p_2 = 3$

		$n = 35$			$n = 50$		
		$\hat{d}$			$\hat{d}$		
ρ	Δ	RL	SL	PSL	RL	SL	PSL
0.90	0.0	4.946	1.456	1.627	3.139	1.198	1.490
	0.4	3.125	1.374	1.439	2.001	1.267	1.319
	0.6	2.115	1.259	1.303	1.313	1.172	1.197
	0.8	1.453	1.196	1.201	0.879	1.111	1.119
	1.0	1.035	1.136	1.137	0.614	1.078	1.078
	2.0	0.303	1.037	1.037	0.173	1.019	1.019
	3.0	0.139	1.017	1.017	0.078	1.008	1.008
	4.0	0.079	1.010	1.010	0.044	1.005	1.005
0.93	0.0	5.627	1.532	1.661	3.365	1.184	1.493
	0.4	3.946	1.450	1.512	2.469	1.285	1.383
	0.6	2.836	1.350	1.385	1.756	1.213	1.269
	0.8	2.030	1.264	1.275	1.238	1.156	1.175
	1.0	1.485	1.192	1.195	0.894	1.109	1.115
	2.0	0.456	1.053	1.053	0.264	1.029	1.029
	3.0	0.211	1.024	1.024	0.121	1.012	1.012
	4.0	0.121	1.014	1.014	0.068	1.007	1.007
0.95	0.0	5.764	1.464	1.595	3.534	1.386	1.507
	0.4	4.418	1.417	1.501	2.853	1.295	1.433
	0.6	3.381	1.343	1.408	2.178	1.245	1.335
	0.8	2.540	1.223	1.317	1.619	1.176	1.243
	1.0	1.922	1.232	1.241	1.210	1.158	1.172
	2.0	0.632	1.070	1.070	0.382	1.044	1.044
	3.0	0.298	1.031	1.031	0.177	1.019	1.019
	4.0	0.171	1.018	1.018	0.101	1.010	1.010

Table 2: The $SREs$ of the proposed estimators with respect to $\hat{\boldsymbol{\beta}}^{UL}$ for $p_2 = 5$

		$n = 35$			$n = 50$		
		$\hat{d}$			$\hat{d}$		
ρ	Δ	RL	SL	PSL	RL	SL	PSL
0.90	0.0	5.661	2.025	2.450	5.035	1.871	2.440
	0.4	3.374	1.787	1.960	2.684	1.715	1.829
	0.6	2.312	1.588	1.650	1.667	1.433	1.510
	0.8	1.614	1.425	1.437	1.086	1.310	1.322
	1.0	1.166	1.301	1.303	0.749	1.216	1.216
	2.0	0.356	1.075	1.075	0.208	1.057	1.057
	3.0	0.166	1.030	1.030	0.094	1.026	1.026
	4.0	0.095	1.015	1.015	0.053	1.014	1.014
0.93	0.0	5.889	2.028	2.461	5.282	1.803	2.465
	0.4	3.911	1.892	2.076	3.274	1.833	1.977
	0.6	2.845	1.641	1.786	2.194	1.577	1.662
	0.8	2.073	1.525	1.560	1.498	1.409	1.443
	1.0	1.542	1.393	1.401	1.063	1.302	1.307
	2.0	0.498	1.107	1.107	0.310	1.084	1.084
	3.0	0.235	1.044	1.044	0.142	1.038	1.038
	4.0	0.136	1.022	1.022	0.081	1.021	1.021
0.95	0.0	6.098	2.071	2.340	5.518	1.987	2.457
	0.4	4.379	1.860	2.046	3.844	1.956	2.105
	0.6	3.354	1.717	1.819	2.762	1.702	1.820
	0.8	2.544	1.578	1.622	1.978	1.495	1.586
	1.0	1.949	1.457	1.472	1.448	1.401	1.420
	2.0	0.672	1.148	1.148	0.447	1.118	1.118
	3.0	0.323	1.065	1.065	0.207	1.052	1.052
	4.0	0.188	1.035	1.035	0.118	1.029	1.029

Table 3: The $SREs$ of the proposed estimators with respect to $\hat{\boldsymbol{\beta}}^{UL}$ for $p_2 = 7$

		$n = 35$			$n = 50$		
		$\hat{d}$			$\hat{d}$		
ρ	Δ	RL	SL	PSL	RL	SL	PSL
0.90	0.0	8.802	2.783	3.203	6.337	2.661	3.269
	0.4	5.482	2.470	2.730	3.680	2.202	2.414
	0.6	3.751	2.156	2.331	2.359	1.860	1.920
	0.8	2.605	1.910	1.992	1.563	1.596	1.609
	1.0	1.871	1.714	1.741	1.088	1.423	1.424
	2.0	0.560	1.238	1.238	0.306	1.120	1.120
	3.0	0.259	1.111	1.111	0.139	1.056	1.056
	4.0	0.148	1.064	1.064	0.079	1.033	1.033
0.93	0.0	9.044	2.756	3.091	7.187	2.770	3.305
	0.4	6.285	2.550	2.822	4.680	2.385	2.621
	0.6	4.593	2.326	2.509	3.188	2.052	2.149
	0.8	3.342	2.096	2.201	2.196	1.773	1.803
	1.0	2.478	1.894	1.944	1.565	1.567	1.573
	2.0	0.788	1.333	1.333	0.458	1.167	1.167
	3.0	0.369	1.156	1.156	0.210	1.077	1.077
	4.0	0.212	1.089	1.089	0.119	1.044	1.044
0.95	0.0	9.157	2.638	2.878	1.044	2.779	3.216
	0.4	6.822	2.490	2.694	5.228	2.466	2.709
	0.6	5.274	2.315	2.474	3.804	2.178	2.299
	0.8	4.018	2.121	2.238	2.745	1.899	1.960
	1.0	3.083	1.944	2.020	2.019	1.691	1.712
	2.0	1.058	1.405	1.406	0.626	1.223	1.223
	3.0	0.507	1.203	1.203	0.291	1.105	1.105
	4.0	0.294	1.119	1.119	0.166	1.061	1.061