Ridge-Penalized Zero-Inflated Probit Bell model for multicollinearity in count data

Essoham Ali, Adewale F Lukman

Ridge-Penalized Zero-Inflated Probit Bell model for multicollinearity in count data

Essoham ALI[1,2*] and Adewale F. LUKMAN[3†]

[1]Institut de Mathématiques Appliquées, UCO, 49000, Angers, France.
[2]Univ Bretagne Sud, CNRS UMR 6205, LMBA, Vannes, France.
[3]Department of Mathematics and Statistics, University of North Dakota, Grand Forks, North Dakota, 58202 USA.

*Corresponding author(s). E-mail(s): essoham.ali@univ-ubs.fr ;
Contributing authors: adewale.lukman@und.edu;
†These authors contributed equally to this work.

Abstract

This article introduces a ridge estimator within the Zero-Inflated Probit Bell (ZIPBell) regression model, developed specifically to handle count data characterized by excess zeros and multicollinearity among predictor variables. By incorporating ridge penalization into the ZIPBell framework, we provide a methodology that stabilizes parameter estimates by reducing variance and mitigating multicollinearity effects without excluding correlated predictors. A numerical study and an empirical application illustrate the robustness of this approach across varying levels of multicollinearity and data sparsity, presenting a reliable tool for analyzing complex count data with structural zeros and correlated predictors.

Keywords: Count data, Zero-Inflated Probit Bell model, Ridge regression, Multicollinearity, Penalized estimation

1 Introduction

Statistical modeling with correlated predictors, or multicollinearity, presents a long-standing challenge, particularly in complex count data. In traditional regression analysis, multicollinearity refers to a situation where the predictive variables are highly correlated, leading to unstable coefficient estimates and increased standard errors [18]. Ridge regression, introduced by [8], is a commonly used technique to address this issue by introducing a penalty term that reduces the estimated variance at the cost of a slight bias. This trade-off often results in more reliable and interpretable models, especially when traditional methods fail due to multicollinearity.

Zero-inflated (ZI) models, initially introduced by [10] were developed to handle datasets with an abundance of zeros, where classical count models such as Poisson and negative binomial models fail to fit adequately. These models assume two processes: one governing the occurrence of zeros, and the other generating non-zero counts. The framework of zero-inflated models has been widely applied across various fields, from health sciences to ecology, for studying rare events, equipment failures, and medical conditions characterized by zero counts [19].

Count datasets often exhibit inflated zeros, which can be addressed using various approaches. The Zero-Inflated Bell (ZIBell) distribution has emerged as a suitable alternative to the commonly used Zero-Inflated Poisson (ZIP) distribution for handling this issue. Recognizing the advantages of the ZIBell

distribution, [12] proposed a new regression model based on the ZIBell distribution, providing a comprehensive examination of its properties. More recently, [2] offered an in-depth summary of the asymptotic properties of the ZIBell model. Building on this foundation, [3] introduced an innovative extension—the Zero-Inflated Probit Bell (ZIPBell) model. This new model refines the zero-inflated approach by incorporating a probit link function to effectively handle binary outcomes, thereby improving predictive accuracy in datasets with a substantial proportion of zeros.

However, while these advances address overdispersion and zero inflation, the challenge of multicollinearity among predictors remains unresolved and can complicate parameter estimation. One of the key limitations in zero-inflated models affected by multicollinearity is the instability of the maximum likelihood estimator (MLE). When predictor variables are highly correlated, MLE often struggles to converge or yields unreliable estimates, undermining model accuracy and interpretability. These limitations have spurred researchers to explore alternative estimation techniques, such as ridge regression, which introduces a penalty term to stabilize coefficient estimates by reducing variance at the expense of some bias [8]. This trade-off can improve model reliability and interpretability, particularly where traditional MLE approaches falter under multicollinearity.

The ridge regression, originally introduced by [8], offers a robust solution for dealing with multicollinearity. By introducing a penalty term on the regression coefficients, ridge regression shrinks the estimates towards zero, decreasing variance and minimizing the impact of collinearity without eliminating correlated predictors. Although initially developed for linear regression contexts, ridge regression has shown considerable success when extended to other regression frameworks, including logistic and count data models ([11]; [16];[17]; [9]; [14]; [1]; [15]; [4]; [5]). For example, [16] adapted ridge regression to Poisson models,[9] applied it to Zero-Inflated Poisson (ZIP) models, and [1] further extended it to the Zero-Inflated Bell (ZIBell) regression model. [4] applied ridge regression to Bell models to further address overdispersion in count data.

The integration of ridge regression into zero-inflated models is relatively recent, with researchers beginning to recognize its potential to enhance estimation stability in theoretical and applied settings [7]. [7] and [6] demonstrated that a penalized likelihood approach can effectively stabilize parameter estimates in complex, high-dimensional datasets, encouraging further exploration. Building on this foundation, our study seeks to extend ridge penalization to the Zero-Inflated Probit Bell (ZIPBell) model. By integrating ridge regression within the ZIPBell framework, we provide theoretical justifications and empirical validation to showcase its effectiveness. This approach aims to counteract the negative effects of multicollinearity while preserving the interpretative power of zero-inflated models.

This extension contributes to the literature on penalized estimation for count data but also broadens the applicability of ridge regression in zero-inflated contexts, where excess zeros and predictor correlations are common. Our proposed model thus addresses the dual challenges of multicollinearity and zero inflation, bridging an important gap between theoretical developments and practical applications. As robust regularization techniques are increasingly needed in advanced count data models, this study opens pathways for further innovations in penalized regression methods.

The article's structure is as follows: Section 2 offers a comprehensive review of the literature on ridge regression, zero-inflated models, and Bell distributions, followed by a detailed presentation of the methodology and formulation of the ridge-Penalized Zero-Inflated Probit Bell (RP-ZIPB) model. Section 3 includes a simulation study to evaluate the model's performance under varying levels of multicollinearity and data sparsity. It also demonstrates the model's application to a real-world dataset, emphasizing its practical relevance and implications. Finally, Section 4 discusses the model's strengths, limitations, and avenues for future research.

2 Preliminaries

In this section, we present the Zero-Inflated Probit Bell (ZIPBell) regression model as provided by [3].

Definition 2.1 (Zero-inflated Probit Bell model):

The random variable Y is said to have zero-inflated Probit Bell distribution, denoted by $Y \sim \text{ZIPBell}(\pi, \phi)$ if The general formula of the ZIPBell model has the following form:

$$P(Y=y|x,s)=\pi I(y=0)+(1-\pi)\exp\left(1-e^{W(\phi)}\right)\frac{W(\phi)^y B_y}{y!}, \tag{1}$$

for $y=0,1,2,\cdots,W(.)$ is the Lambert function, and B_y are the Bell numbers.

When risk factors are available, the mixing probability π_i is usually modelled by a probit regression: probit$(\pi_i)=F(\beta^\top \mathbf{S}_i)$ where F denotes the cumulative distribution function (CDF) of the standard normal distribution, $\mathcal{N}(0,1)$ and ϕ_i is classically modelled as $\phi_i(\beta)=\exp(\alpha^\top \mathbf{X}_i)$. Vectors $\alpha=(\alpha_1,\ldots,\alpha_p)^\top\in\mathbb{R}^p$ and $\beta=(\beta_1,\ldots,\beta_q)^\top\in\mathbb{R}^q$ are unknown regression parameters. Let $J_i=1\{Y_i=0\}$ and $\bar{J}_i=1-J_i$. Suppose that we observe n independent vectors $(Y_i,\mathbf{S}_i,\mathbf{X}_i)$, $i=1,\ldots,n$. Let $\Phi:=(\alpha^\top,\beta^\top)^\top$ denote the set of all unknown parameters. Then, the likelihood function of Φ is

$$L_n(\Phi)=\prod_{i=1}^{n}\left[\pi_i+(1-\pi_i)\exp(1-e^{W(\phi)})\right]^{I(Y_i=0)}\left[(1-\pi_i)\exp(1-e^{W(\phi)})\frac{W(\phi)^{Y_i}B_y}{Y_i!}\right]^{I(Y_i>0)}. \tag{2}$$

Using (2) and some algebra, the loglikelihood $\ell_n(\Phi)=\log Ln(\Phi)$ can be written as :

$$\begin{aligned}\ell(\Phi)\;=\;&\sum_{i=1}^{n}\Bigg\{J_i\log\left[F(\beta^\top s_i)+(1-F(\beta^\top s_i))\exp(1-e^{W(e^{\alpha^\top x_i})})\right]\\&+\bar{J}_i\left[Y_i\log(W(e^{\alpha^\top x_i}))+\log(1-F(\beta^\top s_i))-e^{W(e^{\alpha^\top x_i})}\right]\Bigg\}.\end{aligned} \tag{3}$$

The maximum likelihood estimation (MLE) algorithm can optimize $\ell(\Phi)$ to obtain $\widehat{\Phi}$, the parameter estimate of $\Phi=(\alpha^\top,\beta^\top)^\top$. This optimization is represented as:

$$\widehat{\Phi}_M=\mathbf{argmax}\,\ell(\Phi). \tag{4}$$

2.1 Ridge-Penalized Zero-Inflated Probit Bell (RP-ZIPB) model

When multicollinearity among predictors is problematic, ridge estimation offers a practical remedy, particularly within zero-inflated count models like the Zero-Inflated Probit Bell (ZIPBell) model. Ridge estimators, introduced by [8], add a penalty term to the log-likelihood function to regularize parameter estimates, which helps to stabilize them and reduce variance in the presence of highly correlated variables. Given the ZIPBell model's log-likelihood function in Equation (3), we define the ridge-penalized log-likelihood as follows:

$$\begin{aligned}\ell_{\text{ridge}}(\Phi)&=\ell(\Phi)-\lambda\|\Phi\|^2\\&=\sum_{i=1}^{n}\Bigg\{J_i\log\left[F(\beta^\top s_i)+(1-F(\beta^\top s_i))\exp(1-e^{W(e^{\alpha^\top x_i})})\right]\\&\quad+\bar{J}_i\left[Y_i\log(W(e^{\alpha^\top x_i}))+\log(1-F(\beta^\top s_i))-e^{W(e^{\alpha^\top x_i})}\right]\Bigg\}-\lambda\|\Phi\|^2,\end{aligned} \tag{5}$$

where $\|\Phi\|^2=\alpha^\top\alpha+\beta^\top\beta$ represents the squared Euclidean norm of the parameter vector Φ, and $\lambda>0$ is the ridge penalty parameter controlling the degree of shrinkage.

Maximizing the ridge-penalized log-likelihood function $\ell_{\text{ridge}}(\Phi)$ requires solving for Φ while minimizing the penalty $\lambda\|\Phi\|^2$. When $\lambda=0$, this reverts to standard maximum likelihood estimation (MLE) as described by [3]. Increasing λ results in greater shrinkage of the estimates, mitigating issues of multicollinearity by pulling coefficients toward zero.

To find the ridge estimates $\hat{\Phi}_{\text{ridge}}$, we solve the first-order conditions of the penalized log-likelihood:

$$\frac{\partial \ell_{\text{ridge}}(\Phi)}{\partial \Phi} = \frac{\partial \ell(\Phi)}{\partial \Phi} - 2\lambda\Phi = 0, \tag{6}$$

where $\frac{\partial \ell(\Phi)}{\partial \Phi}$ is the gradient of the log-likelihood function, and $2\lambda\Phi$ is the derivative of the penalty term. By iteratively updating estimates using a modified Newton-Raphson or Fisher scoring algorithm, we achieve convergence to $\hat{\Phi}_{\text{ridge}}$. Each iteration involves:

$$\Phi^{t+1} = \Phi^t - (H_{\text{ridge}}(\Phi^t))^{-1} \nabla \ell_{\text{ridge}}(\Phi^t) \tag{7}$$

where $\nabla \ell_{\text{ridge}}(\Phi^t)$ is the gradient vector of the penalized log-likelihood evaluated at the current estimate Φ^t, and $H_{\text{ridge}}(\Phi^t)$ is the Hessian matrix of the log-likelihood at Φ^t. The algorithm for the ridge regression estimator converges when:

$$\|\boldsymbol{\Phi}^{(t+1)} - \boldsymbol{\Phi}^{(t)}\|_2 < \epsilon, \tag{8}$$

where: $\epsilon > 0$ is a pre-specified tolerance level (e.g., 10^{-6}).

Alternatively, convergence can also be determined based on changes in the penalized log-likelihood:

$$\left|\ell_{\text{ridge}}(\boldsymbol{\Phi}^{(t+1)}) - \ell_{\text{ridge}}(\boldsymbol{\Phi}^{(t)})\right| < \delta, \tag{9}$$

where: $\delta > 0$ is a small value indicating negligible improvement. For models where the log-likelihood $\ell(\boldsymbol{\theta})$ is approximated quadratically, the ridge estimate can be directly computed in closed form:

$$\hat{\boldsymbol{\Phi}}_{\text{ridge}} = (\mathbf{H} + k\mathbf{I})^{-1} \nabla \ell, \tag{10}$$

where: $\mathbf{H} = \frac{\partial^2 \ell(\boldsymbol{\Phi})}{\partial \boldsymbol{\Phi} \partial \boldsymbol{\Phi}^\top}$ is the unpenalized Hessian, $\nabla \ell = \frac{\partial \ell(\boldsymbol{\Phi})}{\partial \boldsymbol{\Phi}}$ is the unpenalized gradient, and $k > 0$ is the ridge penalty parameter. We adopted the ridge parameters by [1]. The ridge parameters are defined as follows:

$$k_1 = \frac{p}{\sum_{j=1}^{p} \hat{\Phi}_{Mj}^2}, \tag{11}$$

$$k_2 = \frac{1}{\min\left(\hat{\Phi}_{Mj}^2\right)}, \tag{12}$$

$$k_3 = \text{median}\left(\frac{1}{\hat{\Phi}_{Mj}^2}\right), \tag{13}$$

$$k_4 = \left(\prod_{j=1}^{p} \hat{\Phi}_{Mj}^2\right)^{\frac{1}{p}}, \tag{14}$$

$$k_5 = \sqrt{\frac{p}{\sum_{j=1}^{p} \hat{\Phi}_{Mj}^2}}. \tag{15}$$

where: p is the number of predictors, $\hat{\Phi}_{Mj}$ represents the maximum likelihood estimate (MLE) of the j-th parameter.

This closed-form solution is not generally used for estimation but is critical for deriving analytical properties, such as bias, variance, and SMSE. Using the closed-form approximation, we derive key theoretical metrics. The ridge estimate introduces bias due to the penalty term:

$$\text{Bias}(\hat{\boldsymbol{\Phi}}_{\text{ridge}}) = -k(\mathbf{H} + k\mathbf{I})^{-1} \boldsymbol{\Phi}_{\boldsymbol{M}}. \tag{16}$$

The variance of $\hat{\mathbf{\Phi}}_{\text{ridge}}$ accounts for both the penalty term and the variability in the data. The variance-covariance matrix is:

$$\text{Var}(\hat{\mathbf{\Phi}}_{\text{ridge}}) = (\mathbf{H} + k\mathbf{I})^{-1}\mathbf{H}(\mathbf{H} + k\mathbf{I})^{-1}. \tag{17}$$

The Mean squared error (MSE) is defined as follows:

$$\text{MSE}(\hat{\Phi}_{\text{ridge}}) = \text{Bias}^2(\hat{\Phi}_{\text{ridge}}) + \text{Var}(\hat{\Phi}_{\text{ridge}}). \tag{18}$$

The scalar Mean Squared Error (SMSE), which is defined as the trace of the covariance matrix of the estimator, is expressed as follows:

$$\text{MSE}(\hat{\mathbf{\Phi}}_{\text{ridge}}) = \sum_{j=1}^{p} \left[\left(\frac{k}{H_j + k} \Phi_j \right)^2 + \frac{H_j}{(H_j + k)^2} \right], \tag{19}$$

where H_j represents the diagonal entries of the Hessian, and p is the total number of parameters.

$$\sqrt{n} \left(\hat{\Phi}_{\text{ridge}} - \Phi_0 \right) \xrightarrow{d} N\left(0, \Sigma_{\text{ridge}}\right) \tag{20}$$

where Σ_{ridge} is the variance-covariance matrix of the ridge estimator defined in equation (17).

2.2 Pseudo-Code for RI-ZIPBRM

Algorithm 1 RP-ZIPB Estimation Procedure

Input: Dataset with response variable y and predictors $x_1, x_2, \ldots, x_p$.
Output: Ridge-penalized estimates $\hat{\Phi}_{\text{ridge}}$, bias, variance, and MSE.

Step 1: Preprocessing
Normalize predictors and create model matrices X and S.

Step 2: Define the ZIPBell Log-Likelihood
Formulate the log-likelihood $\ell(\Phi)$.

Step 3: Ridge Penalty
Incorporate ridge penalty $\lambda\|\Phi\|^2$ into $\ell(\Phi)$.

Step 4: Optimization
Solve the penalized log-likelihood using iterative optimization:

$$\Phi^{t+1} = \Phi^t - (H_{\text{ridge}}(\Phi^t))^{-1} \nabla \ell_{\text{ridge}}(\Phi^t).$$

Step 5: Derive Theoretical Properties
Use the closed-form solution to compute:

- Bias: $-\lambda(\mathbf{H} + \lambda\mathbf{I})^{-1}\Phi$.
- Variance: $(\mathbf{H} + \lambda\mathbf{I})^{-1}\mathbf{H}(\mathbf{H} + \lambda\mathbf{I})^{-1}$.
- SMSE: trace(MSE).

Step 6: Validate and Summarize Results
Assess multicollinearity and model stability.
Summarize estimates, diagnostics, and visualizations.

3 Empirical studies

3.1 Simulation

In this section, we evaluate the performance of the Ridge estimator for the ZIPBell model and compare it with the performance of the Maximum Likelihood Estimation (MLE). To achieve this, the count data Y_i are simulated based on the ZIPBell regression model, which is generated using the following model structure:

$$P(Y = y \mid X, S) = \begin{cases} F(S^\top\beta) + \left[1 - F(S^\top\beta)\right] \exp(1 - e^{W(\phi)}) & \text{if } y = 0, \\ \left[1 - F(S^\top\beta)\right] \exp(1 - e^{W(\phi)}) \frac{W(\phi)^y B_y}{y!} & \text{if } y > 0, \end{cases} \tag{21}$$

where: $F(S^\top\beta)$ is the cumulative distribution function (CDF) of the standard normal distribution (probit link), $W(\phi)$ is the Lambert W function, B_y are the Bell numbers, and $\phi = \exp(X^\top\alpha)$.

The explanatory variables are $X = (X_1, X_2, \ldots, X_p)$ and $S = (S_1, S_2, \ldots, S_q)$. For simplicity, the covariates $S_1 = X_2$ and $S_2 = X_5$ were used to share components across models [12]. Correlated explanatory variables are generated using the following equation:

$$x_{ij} = \sqrt{1 - \rho^2} m_{ij} + \rho m_{ip}, \tag{22}$$

for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$. Here, m_{ij} represents pseudo-random numbers drawn from the standard normal distribution, and ρ indicates the correlation between the explanatory variables, with values of $\rho = 0.6, 0.9, 0.95, 0.99$. For each simulation, the sample sizes considered are $n = 50, 100, 150, 300, 500, 1000$. The number of explanatory variables p takes values of $p = 5$ and $p = 8$ to examine how the performance of the estimators changes as the number of predictors increases. Each simulation is repeated 1000 times for every combination of the specified parameters. The performance of the different estimation methods is evaluated using the Mean Squared Error (MSE) as the comparison metric.

$$\text{MSE}(\hat{\Phi}) = \frac{1}{1000} \sum_{r=1}^{1000} (\hat{\Phi}_r - \Phi)'(\hat{\Phi}_r - \Phi)$$

where $\hat{\Phi}_r$ denotes the estimated vector of the true parameter vector Φ in rth replication.

Tables 1 and 2 in the document present simulated mean square error (MSE) values for different estimators in the Zero-Inflated Probit Bell (ZIPBell) model with ridge penalization, used to handle multicollinearity in count data. We compare the classical maximum likelihood estimation (MLE) with several versions of ridge estimators, denoted as $\hat{k}_1, \hat{k}_2, \ldots, \hat{k}_5$, according to the correlation levels (ρ) and sample size (n). The values of ρ range from 0.60 to 0.99, illustrating increasing levels of multicollinearity among predictors. Sample sizes increase from 50 to 1000 observations, enabling the evaluation of estimator performance relative to the amount of data available.

In Table 1, where $p = 5$ (number of predictors), MSE values for each estimator are measured under different combinations of ρ and n. The MLE shows higher MSE values than ridge estimators, particularly when multicollinearity is strong ($\rho = 0.90, 0.95$, or 0.99). This indicates that ridge estimators effectively reduce prediction error under severe multicollinearity, as expected by their design. It is observed that increasing n consistently reduces MSE for all estimators, indicating improved performance with more data. This trend is clearly illustrated in Figure 1. Ridge estimators tend to converge to more stable values and outperform MLE more notably in small samples, where multicollinearity particularly affects MLE variance.

As ρ increases, MLE performance deteriorates faster than ridge estimators. Estimators $\hat{k}_3$ and $\hat{k}_4$ stand out with lower MSE values in cases of high multicollinearity (e.g., $\rho = 0.95$ and $\rho = 0.99$), suggesting they offer a more suitable penalty for extreme correlation levels. In Table 2, where $p = 8$, the same trends observed in Table 1 are confirmed. As anticipated, the MSE values increase with higher levels of multicollinearity (ρ), highlighting the adverse impact of multicollinearity on estimator performance. This trend is depicted in Figure 2.

Compared to the results for $p = 5$, the increase in the number of predictors (from 5 to 8) amplifies the effect of multicollinearity, particularly for non-penalized estimators. MLE MSE values are significantly higher under strong multicollinearity, as for $\rho = 0.95$ and $n = 50$, highlighting the increased instability of unpenalized estimates in models with richer variable sets. Ridge estimators, particularly $\hat{k}_3$ and $\hat{k}_4$,

continue to outperform MLE by reducing MSE values. At $\rho = 0.99$, ridge estimators have noticeably lower MSE values, confirming the advantage of ridge penalization in situations of extreme predictor correlation.

The results of both tables demonstrate that ridge estimators provide more stable estimates than MLE, especially when multicollinearity is high and the sample size is small. Using different penalty coefficients λ in ridge estimators allows adjusting the regularization degree according to the intensity of multicollinearity and sample size, optimizing the balance between bias and variance. These tables confirm that integrating ridge penalization in the ZIPBell model effectively addresses the challenges of multicollinearity in count data. Ridge estimators reduce prediction error (MSE), particularly in situations of high multicollinearity or small samples, thus offering a robust alternative to maximum likelihood estimation.

Table 1: Simulated MSE values of the estimators when $p = 5$. Note that all results are based on $N = 1000$ simulated samples.

			RP-ZIPB						RP-ZIPB				
ρ	n	MLE	$\widehat{k}_1$	$\widehat{k}_2$	$\widehat{k}_3$	$\widehat{k}_4$	$\widehat{k}_5$	MLE	$\widehat{k}_1$	$\widehat{k}_2$	$\widehat{k}_3$	$\widehat{k}_4$	$\widehat{k}_5$
0.60	50	1.0787	0.5744	0.4930	0.4263	0.4402	0.5314	1.7680	0.8303	0.6243	0.4699	0.5119	0.7024
	100	0.8081	0.4521	0.3742	0.3454	0.3498	0.4365	0.7489	0.5190	0.4507	0.4143	0.3974	0.3929
	150	0.6478	0.4302	0.3814	0.3428	0.3495	0.4463	0.4139	0.3791	0.3790	0.3890	0.3703	0.2487
	300	0.5476	0.3498	0.3313	0.3216	0.3231	0.3851	0.2148	0.3449	0.3739	0.3908	0.3816	0.1956
	500	0.5509	0.3424	0.3289	0.3214	0.3224	0.3842	0.1595	0.3558	0.3779	0.3937	0.3884	0.2161
	1000	0.5294	0.3294	0.3239	0.3209	0.3213	0.3666	0.1166	0.3672	0.3839	0.3956	0.3932	0.2343
0.90	50	1.5990	1.0341	0.8861	0.6612	0.6952	0.8648	3.0285	1.6844	1.4025	1.0320	1.1289	1.4296
	100	1.0996	0.6895	0.5873	0.4247	0.4380	0.5902	1.4466	0.8960	0.7811	0.5609	0.5554	0.6651
	150	0.8655	0.5869	0.5057	0.3882	0.4007	0.5306	0.7131	0.5604	0.5243	0.4169	0.3914	0.3910
	300	0.7836	0.4909	0.4287	0.3422	0.3482	0.4869	0.4595	0.4105	0.4102	0.3926	0.3676	0.2579
	500	0.7105	0.4185	0.3763	0.3245	0.3283	0.4489	0.3504	0.3727	0.3941	0.3967	0.3795	0.2268
	1000	0.6860	0.3793	0.3500	0.3219	0.3240	0.4386	0.2609	0.3435	0.3641	0.3923	0.3841	0.2236
0.95	50	2.8883	2.0835	1.8488	1.4714	1.5018	1.6622	4.9427	3.1970	2.9245	2.3698	2.4967	2.6903
	100	1.6284	1.1380	1.0066	0.7183	0.7340	0.8933	1.9993	1.4920	1.3459	0.9506	0.9616	1.1312
	150	1.3376	0.9348	0.8405	0.5639	0.5784	0.7385	1.5230	1.1790	1.0870	0.6559	0.6721	0.8418
	300	0.9827	0.6908	0.6004	0.3678	0.3900	0.6176	0.7428	0.6078	0.5834	0.3894	0.3714	0.4258
	500	0.7985	0.5325	0.4610	0.3347	0.3463	0.5248	0.4539	0.4225	0.4257	0.3767	0.3543	0.2758
	1000	0.9018	0.5239	0.4648	0.3246	0.3307	0.5385	0.4951	0.4355	0.4366	0.3860	0.3706	0.2894
0.99	50	13.2242	9.4606	8.8688	7.8065	7.9824	7.7133	20.6767	15.3167	15.0525	13.7919	14.1487	13.7892
	100	6.0860	5.3068	4.8552	3.8485	4.0143	3.9744	9.9855	8.4337	8.2387	6.7842	7.1452	6.9088
	150	4.4873	3.9861	3.6544	2.8801	2.9462	2.9054	6.3439	5.9124	5.7441	4.5689	4.6944	4.3844
	300	2.7675	2.2077	2.0543	1.2997	1.3447	1.6333	3.2721	2.8768	2.7540	1.6972	1.7207	1.9493
	500	2.3539	2.0340	1.9382	1.0427	1.0758	1.3987	2.5663	2.3868	2.2998	1.1713	1.1664	1.4462
	1000	1.8323	1.5385	1.4472	0.6326	0.6638	1.0443	1.8899	1.7322	1.6583	0.7117	0.7059	1.0098

Table 2: Simulated MSE values of the estimators when $p = 8$. Note that all results are based on $N = 1000$ simulated samples.

			RP-ZIPB						RP-ZIPB				
ρ	n	MLE	$\widehat{k}_1$	$\widehat{k}_2$	$\widehat{k}_3$	$\widehat{k}_4$	$\widehat{k}_5$	MLE	$\widehat{k}_1$	$\widehat{k}_2$	$\widehat{k}_3$	$\widehat{k}_4$	$\widehat{k}_5$
0.60	50	0.7722	0.2992	0.2604	0.2501	0.2521	0.2972	1.7395	0.5221	0.4912	0.4631	0.4680	0.5638
	100	0.7082	0.2760	0.2530	0.2490	0.2502	0.2868	0.7422	0.4058	0.4054	0.4330	0.4031	0.3511
	150	0.4975	0.2657	0.2500	0.2476	0.2483	0.2778	0.4798	0.3508	0.4141	0.4368	0.4139	0.2635
	300	0.4153	0.2498	0.2464	0.2462	0.2463	0.2609	0.2080	0.3367	0.4267	0.4419	0.4290	0.1807
	500	0.3994	0.2484	0.2464	0.2463	0.2463	0.2575	0.1349	0.3650	0.4345	0.4437	0.4366	0.1884
	1000	0.3710	0.2474	0.2464	0.2463	0.2464	0.2516	0.0763	0.4024	0.4416	0.4460	0.4427	0.2151
0.90	50	1.5882	0.5972	0.4508	0.3535	0.3702	0.5335	4.1336	1.3891	1.1310	0.8370	0.9532	1.3070
	100	0.9061	0.4814	0.3578	0.2906	0.3001	0.4411	1.7511	0.9407	0.7085	0.5472	0.5688	0.8408
	150	0.7557	0.3747	0.2916	0.2631	0.2655	0.3700	1.0132	0.6345	0.5091	0.4675	0.4512	0.5379
	300	0.5263	0.3050	0.2594	0.2512	0.2522	0.2522	0.4367	0.4191	0.4366	0.4460	0.4250	0.2858
	500	0.4340	0.2726	0.2505	0.2475	0.2477	0.2792	0.2350	0.3627	0.4225	0.4381	0.4248	0.2067
	1000	0.3954	0.2549	0.2474	0.2466	0.2466	0.2614	0.1289	0.3754	0.4333	0.4426	0.4351	0.2069
0.95	50	2.5849	0.8300	0.5753	0.5753	0.6231	0.8913	7.5638	2.6572	2.3570	1.6557	1.9565	2.4798
	100	1.5044	0.8619	0.6482	0.4538	0.4741	0.6891	3.0252	1.8124	1.4181	0.9663	1.0740	1.5744
	150	1.0083	0.6121	0.4336	0.3237	0.3364	0.5307	1.7175	1.1527	0.8320	0.6111	0.6358	0.9809
	300	0.6779	0.4024	0.3047	0.2678	0.2706	0.3736	0.7346	0.5933	0.5069	0.4668	0.4456	0.4488
	500	0.5820	0.3371	0.2801	0.2618	0.2613	0.3242	0.4633	0.4336	0.4506	0.4517	0.4290	0.2832
	1000	0.4573	0.2915	0.2608	0.2499	0.2504	0.2880	0.2492	0.3990	0.4402	0.4415	0.4297	0.2228
0.99	50	9.8399	5.2838	4.9676	3.9744	4.1358	4.2326	28.7262	11.5900	11.3741	9.9768	10.5058	10.8848
	100	5.3619	3.5922	3.1586	2.2535	2.3778	2.6914	13.4107	7.9429	7.4806	5.6483	6.2591	7.0829
	150	1.0083	0.6121	0.4336	0.3237	0.3364	0.5307	1.7175	1.1527	0.8320	0.6111	0.6358	0.9809
	300	2.0868	1.5200	1.2526	0.9095	0.9358	1.0905	3.6030	2.8951	2.4008	1.7481	1.8413	2.2335
	500	1.3746	1.0575	0.8299	0.5317	0.5631	0.8522	1.8647	1.6781	1.3349	0.9223	0.9304	1.2803
	1000	0.7458	0.5212	0.3854	0.3034	0.3155	0.4958	0.7572	0.6847	0.5629	0.5021	0.4802	0.5369

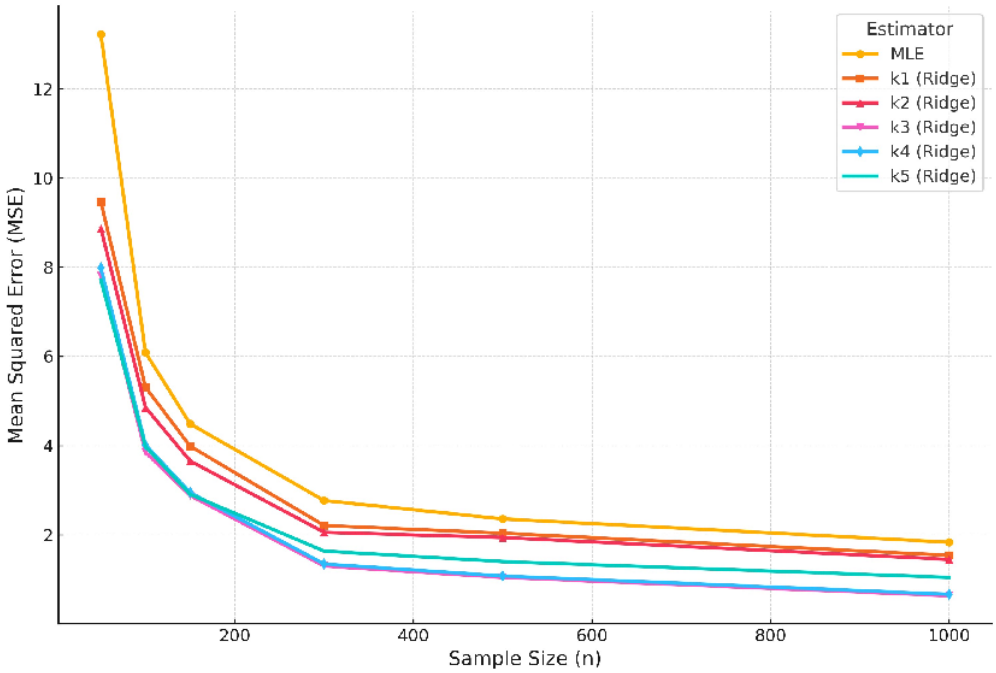

Figure 1: MSE against sample size for $\rho = 0.99$ using the data from Table 1

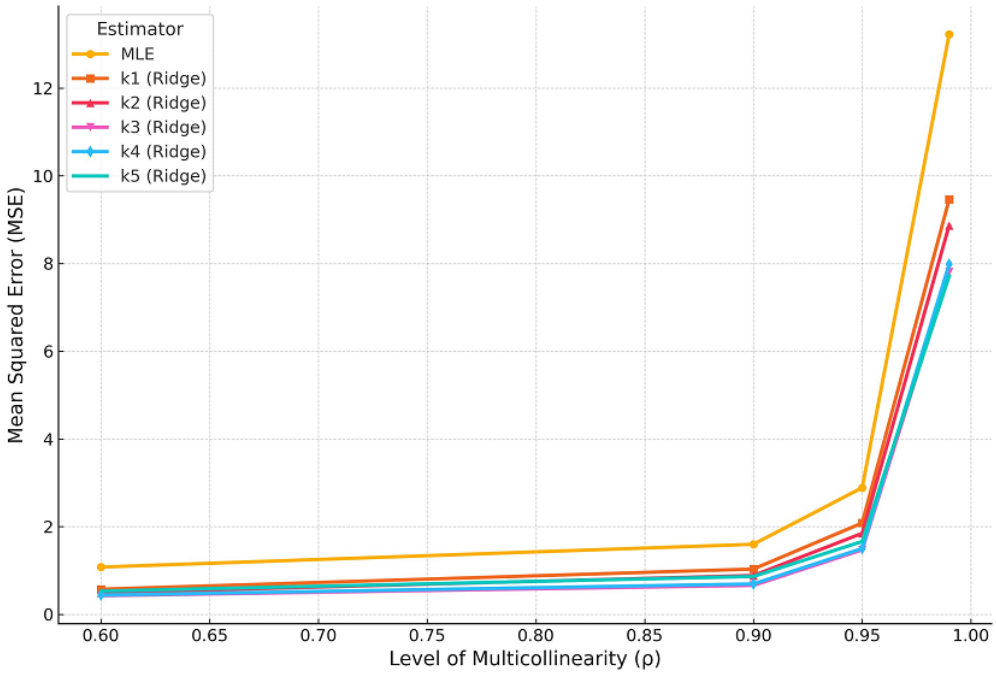

Figure 2: MSE against the level of multicollinearity for $n = 50$ using the data from Table 1

3.2 Real data analysis

In this section, we evaluate the performance of the proposed methods using two real-life datasets, namely the blood transfusion data and the pollutant emissions data.

3.2.1 Application 1: Blood transfusion dataset

The dataset represents the blood transfusions received by 150 randomly selected thalassemia patients in Mosul, Iraq [1]. The following explanatory variables were recorded for each patient: x_1 (age in months), x_2 (duration of thalassemia in months), x_3 (haemoglobin concentration), x_4 (packed cell volume), x_5 (number of blood units), and x_6 (age at the onset of blood transfusion in months). The dataset exhibits a zero-inflation ratio of 0.52, as shown in Figure 3, indicating a substantial proportion of zeros in the response variable. This justifies the suitability of the Zero-Inflated Probit Bell (ZIPBell) model for analyzing the data. The correlation heatmap in Figure 4 revealed significant relationships between certain predictor variables, indicating the presence of multicollinearity. This finding underscores the necessity of employing regularization methods, such as ridge regression, as utilized in this study. To further corroborate these findings, we conducted a Variance Inflation Factor (VIF) analysis to quantify and confirm the degree of multicollinearity among the predictors. The Variance Inflation Factor (VIF) is calculated as $\text{VIF}_j = \frac{1}{1-R_j^2}$, and R_j^2 represents the coefficient of determination from regressing x_j on the remaining explanatory variables. Based on the VIF analysis, variables x_1 (VIF = 37.51) and x_5 (VIF = 35.12) exhibit strong multicollinearity, while x_3 (VIF = 14.25) and x_4 (VIF = 11.13) exhibit moderate multicollinearity. On the other hand, x_2 (VIF = 2.66) and x_6 (VIF = 2.86) have relatively low VIFs, suggesting a weak correlation with the other variables. To overcome multicollinearity, the Ridge method is used in the ZIPBell model. After adjusting the ZI model, the mean squared errors (MSE), standard errors (SE) of the maximum likelihood estimator (MLE), and RP-ZIPB with different ridge parameters are calculated using equations (11) and (15). Table 3 presents the estimated coefficients (β and α), standard errors (SE), and MSE values for different estimators (MLE and Ridge estimators). The Ridge estimators, particularly using k_3 and k_4 as the biasing parameters, outperform MLE in terms of SMSE. This improvement

is reflected by a significant reduction in the SMSE values for β and α, indicating greater stability of the estimated coefficients. For example, for β_1, the ridge estimator with biasing parameter k_3 shows a much lower standard error ($se(k_3) = 0.0002$) than the MLE ($se = 0.1352$).

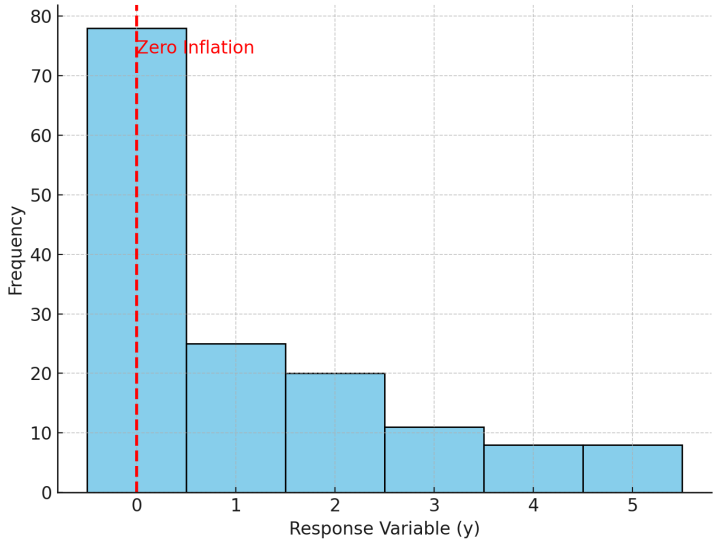

Figure 3: Histogram of the Blood Transfusion Dataset.

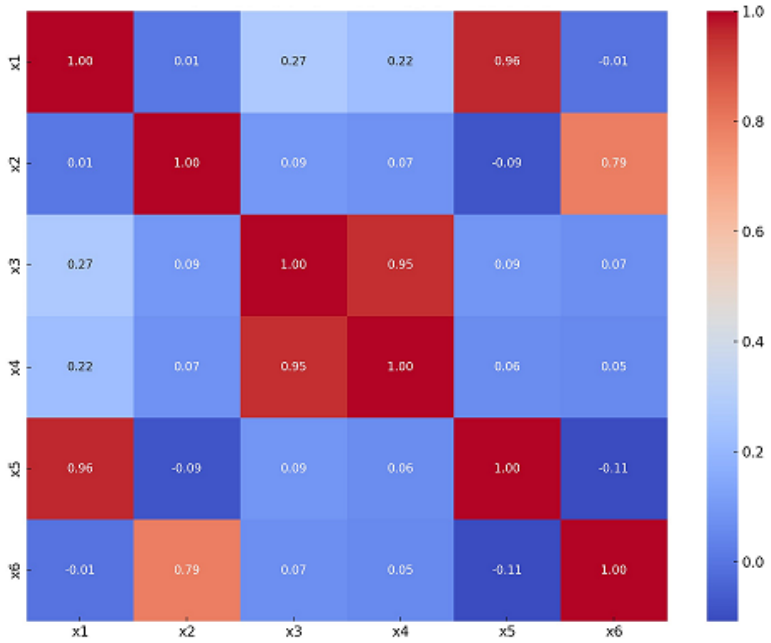

Figure 4: Correlation heatmap of predictors.

3.2.2 Application 2: pollutant emissions dataset

We consider another empirical application to illustrate the advantages of the proposed estimator. This dataset concerns the daily emissions count from a specific pollutant (such as NO_2) measured in an urban area. The sample size of this dataset includes $n = 100$ observations with a response variable and four explanatory variables. The zero inflation reflects days when no detectable emissions were recorded (e.g., days of heavy rain that disperse pollutants). The explanatory variable x_1 represents the average ozone (O_3) concentration, x_2 represents the daily air quality index, and x_3 represents the average daily temperature. The histogram of y is displayed in Figure 5. The dataset exhibits a zero-inflation ratio of 0.53, as shown in Figure 5, indicating a substantial proportion of zeros in the response variable. The correlation heatmap in Figure 6 revealed strong positive correlations among the variables: x_1 and x_2 have a very high correlation ($r = 0.97$), suggesting that emissions represented by these variables are closely related. x_1 and x_3 show a strong positive correlation ($r = 0.88$), though slightly weaker than x_1 and x_2. x_2 and x_3 also have a strong positive correlation ($r = 0.86$). These results indicate that all three variables are interconnected, likely reflecting similar underlying phenomena or processes in pollution emissions. Due to the strong correlations observed among the variables, we further investigated the potential for multicollinearity in the model using the condition index. The condition index for this dataset was calculated to be 254.3948, which strongly indicates the presence of severe multicollinearity among the explanatory variables.

Consequently, we adopted the RP-ZIPB estimation method to address the effect of multicollinearity. Table 4 compares the performance of different estimators (MLE and Ridge) on a dataset of pollutant emissions. The coefficients (β_i and α_i) estimated using Ridge decrease progressively in magnitude as the penalization increases (k_1 to k_5), unlike those obtained by MLE, which remain larger. For example, β_1 decreases from 0.5296 (MLE) to 0.0146 ($\widehat{k}_5$), demonstrating that Ridge reduces the influence of less significant terms. This decrease highlights the effect of regularization, which tends to stabilize estimates by minimizing non-essential contributions.

The standard errors of the estimated coefficients are consistently smaller for Ridge than for MLE, reflecting improved precision. As the penalization increases, the standard errors further decrease, indicating a reduction in uncertainty associated with the estimates. For instance, for β_1, *se* drops from 0.9528 (MLE) to 0.1137 ($\widehat{k}_5$). These results confirm that Ridge regularization, by introducing an additional constraint, enhances the robustness of the estimated coefficients.

Finally, the mean squared errors (MSE) show that Ridge offers better predictive performance compared to MLE. The MSE values for the coefficients (β and α) are significantly reduced with Ridge, decreasing from 83.168 and 109.109 (MLE) to 0.0011 and 0.0007 ($\widehat{k}_5$), respectively. This demonstrates that Ridge regularization is particularly suitable in situations where MLE estimates may be unstable or prone to large errors, while emphasizing that excessive penalization could lead to overly diminished coefficients.

The two empirical studies (blood transfusions and pollutant emissions) demonstrate that the RP-ZIPB model outperforms the classical maximum likelihood estimation (MLE), not only in terms of MSE but also in terms of the stability of the estimated coefficients. These improvements are crucial for decision-making based on complex statistical models.

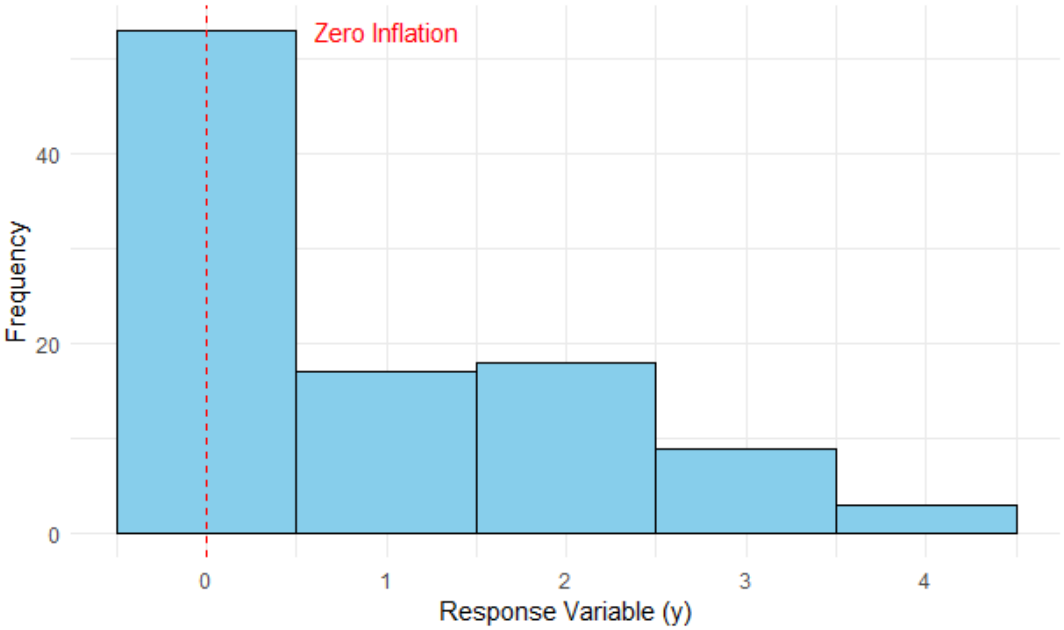

Figure 5: Histogram of the daily number of NO_2

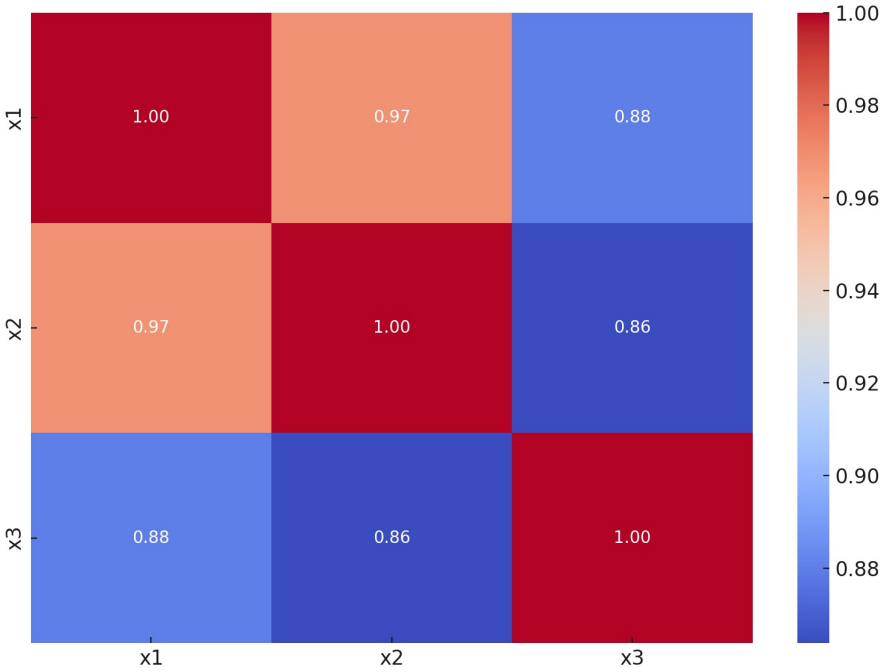

Figure 6: Correlation Heatmap for Pollution Emission Data

Table 3: Estimated coefficients, standard errors, and MSE values for the specified estimators in the fish dataset.

		Ridge estimators										
	MLE	$\widehat{k}_1$	$\widehat{k}_2$	$\widehat{k}_3$	$\widehat{k}_4$	$\widehat{k}_5$	se	se($\widehat{k}_1$)	se($\widehat{k}_2$)	se($\widehat{k}_3$)	se($\widehat{k}_4$)	se($\widehat{k}_5$)
β_1	0.4252	-0.0373	-0.0142	-0.0002	-0.0048	0.0382	0.1352	0.6444	0.0164	0.0002	0.0047	2.2954
β_2	0.0765	0.0960	-0.0422	-0.0007	-0.0177	0.1457	0.6549	0.7684	0.0710	0.0008	0.0219	1.5742
β_3	-0.0100	-0.0126	-0.0075	-0.0001	-0.0028	-0.0092	0.1737	0.8213	0.0212	0.0002	0.0060	2.3859
β_4	0.7418	0.5511	0.1102	0.0014	0.0377	0.6431	0.4566	0.7741	0.0708	0.0008	0.0215	1.6379
β_5	-0.5258	-0.5469	-0.1030	-0.0013	-0.0347	-0.5963	0.5947	0.7940	0.0669	0.0007	0.0202	1.6881
β_6	-0.3612	-0.1676	0.0256	0.0005	0.0120	-0.2346	0.6305	0.7634	0.0687	0.0008	0.0211	1.6125
α_1	-0.0463	-0.0542	-0.0051	0.0000	-0.0014	-0.0607	0.1615	0.7508	0.0196	0.0002	0.0056	2.2597
α_2	-0.9065	-0.5900	-0.1173	-0.0015	-0.0394	-0.7332	0.3983	1.0980	0.0452	0.0005	0.0135	1.8445
α_3	-0.3446	-0.4287	-0.2103	-0.0033	-0.0811	-0.4157	1.2258	0.8269	0.1105	0.0015	0.0386	2.3075
α_4	0.1043	0.0629	0.0048	0.0000	0.0010	0.0778	0.3285	1.0674	0.0397	0.0004	0.0114	1.8039
α_5	2.4730	2.2236	0.5623	0.0074	0.1933	2.3845	1.1498	0.6958	0.1072	0.0014	0.0367	1.8810
α_6	-1.9371	-1.9440	-0.4986	-0.0065	-0.1704	-1.9754	1.0288	0.7876	0.0983	0.0013	0.0331	2.3237
MSE(β)	364.4566	4.1982	0.0459	0.000007	0.0050	24.7864						
MSE(α)	114.0265	12.0328	0.6564	0.000116	0.0787	28.6658						

Table 4: Estimated coefficients, standard errors, and MSE values for the specified estimators in the pollutant emissions dataset.

		Ridge estimators										
	MLE	$\widehat{k}_1$	$\widehat{k}_2$	$\widehat{k}_3$	$\widehat{k}_4$	$\widehat{k}_5$	se	se($\widehat{k}_1$)	se($\widehat{k}_2$)	se($\widehat{k}_3$)	se($\widehat{k}_4$)	se($\widehat{k}_5$)
β_1	0.5296	0.1708	0.1363	0.1085	0.0146	0.0822	0.9528	0.2426	0.1914	0,1510	0.0198	0.1137
β_2	0.0661	-0.0924	-0.0769	-0.0629	-0.0090	-0.0488	0.7954	0.2149	0.1667	0.1300	0.0166	0.0970
β_3	-0.4293	-0.0813	-0.0619	-0.0476	-0.0058	-0.0351	0.4436	0.1559	0.1136	0.0846	0.0094	0.0605
α_1	0.5393	0.1273	0,1003	0,0791	0,0103	0,0595	0,7837	0.2240	0,1721	0,1331	0,0165	0,0984
α_2	0.1733	-0.0638	-0,0528	-0,0430	-0,0061	-0,0332	0,6773	0.2021	0,1535	0,1178	0,0143	0,0865
α_3	-0.4949	-0.0663	-0.0499	-0.0380	-0.0045	-0.0277	0.3713	0.1367	0.0986	0.0728	0.0079	0.0518
MSE(β)	83.168	0.1580	0.0976	0.0607	0.0011	0.0345						
MSE(α)	109.109	0.1303	0.0765	0.0456	0.0007	0.0249						

4 Conclusions

Zero-inflated (ZI) models, first introduced by [10], were specifically designed to address datasets characterized by an excessive number of zeros, where traditional count models such as Poisson and Negative Binomial often fail to provide adequate fit. While classical models assume equidispersion or slight overdispersion, real-world count data often exhibit an excess of zeros, which these models are ill-equipped to handle. This limitation necessitates the development of more flexible alternatives. The Zero-Inflated Bell (ZIBell) distribution has emerged as a robust solution, offering distinct advantages over the Zero-Inflated Poisson (ZIP) distribution by effectively addressing the zero-inflation issue while providing greater flexibility in modeling the count component.

Building on the foundational work of [12] , which introduced the ZIBell regression model and explored its theoretical properties, [2] provided a rigorous examination of the asymptotic properties of the ZIBell framework. Recently, [3] extended this line of research by proposing the Zero-Inflated Probit Bell (ZIPBell) model, incorporating a probit link function for the binary outcome component. This advancement enhances the model's capacity to predict binary zero-inflated structures while maintaining robustness in datasets with a high proportion of zeros.

Despite these innovations, a critical challenge persists in the form of multicollinearity among predictors, which inflates the variance of parameter estimates and undermines the reliability of classical estimation techniques such as MLE. Recognizing this gap, our study introduces the Ridge-Penalized Zero-Inflated Probit Bell (RP-ZIPB) regression model. This enhanced framework combines ridge penalization with the ZIPBell structure to simultaneously address the challenges of multicollinearity, overdispersion, and excess zeros.

Integrating of ridge penalization into the ZIPBell model stabilizes parameter estimates by mitigating multicollinearity's effects while preserving the predictive contributions of correlated variables. Through extensive numerical simulations and empirical applications, the proposed methodology demonstrates superior performance, particularly under scenarios of severe multicollinearity and data sparsity. The results reveal a consistent reduction in estimator variance, improved model fit, and robust predictive capabilities, establishing the Ridge-Penalized ZIPBell model as a versatile and reliable tool for analyzing complex count data. It is obvious that the performance of the proposed method is a function of the adopted biasing parameter.

This work extends the theoretical utility of the ZIPBell framework but also bridges a critical methodological gap in regression analysis for zero-inflated count data with correlated predictors. Future research could explore the integration of adaptive ridge penalties and alternative regularization techniques to enhance further the model's scalability and efficiency in high-dimensional settings. Additionally, comparative analyses with other penalized ZI models could provide deeper insights into the practical advantages of this approach across diverse fields such as public health, econometrics, and environmental science.

Author Contributions

Conceptualization: E. A. and A. F. L.; Methodology: E. A. and A. F. L.; Software: E. A. and A. F. L.; Formal Analysis: E. A. and A. F. L.; Resources: E. A. and A. F. L.; Writing—Original Draft Preparation: E. A. and A. F. L.; Review and Editing: E. A. and A. F. L.; Visualization: E. A. and A. F. L.

Conflict of interest

The authors have no conflicts of interest to declare relevant to this article's content.

Funding

This research received no external funding.

Data Availability Statement

The data will be made available upon request from the corresponding author

References

[1] Algamal, Z. Y., Lukman, A. F., Abonazel, M. R., & Awwad, F. A. Performance of the Ridge and Liu Estimators in the zero-inflated Bell Regression Model. *Journal of Mathematics*, 2022(1), 9503460, 2022.

[2] Ali, E., Diop, M. L., & Diop, A. Statistical Inference in a Zero-Inflated Bell Regression Model. *Mathematical Methods of Statistics*, 31(3), 91-104, 2022.

[3] Ali, E., & Pho, K. H. A novel model for count data: zero-inflated Probit Bell model with applications. *Communications in Statistics - Simulation and Computation*.2024. Available from: https://doi.org/10.1080/03610918.2024.2384574

[4] Amin, M., Akram, M. N., & Majid, A. On the estimation of Bell regression model using ridge estimator. *Communications in Statistics - Simulation and Computation*, 52(3), 854-867,2023.

[5] Bulut, Y. M., Lukman, A. F., Işılar, M., Adewuyi, E. T., & Algamal, Z. Y. Modified ridge estimator in the Bell regression model. *Journal of Inverse and Ill-posed Problems*, 2024. Available from: https://doi.org/10.1515/jiip-2022-0069

[6] Cessie, S. L., and Houwelingen, J. V. "Ridge estimators in logistic regression." *Journal of the Royal Statistical Society Series C: Applied Statistics*, vol. 41, no. 1, pp. 191–201, 1992.

[7] Goeman, J. J., Meijer, R. J., & Chaturvedi, N. Penalized estimation methods for zero-inflated regression models. *Statistical Modelling*, 14(3), 215-237, 2014.

[8] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67, 1970.

[9] Kibria, B. G., Månsson, K., & Shukur, G. Some ridge regression estimators for the zero-inflated Poisson model. *Journal of Applied Statistics*, 40(4), 721-735, 2013.

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

[11] Lee, A. H., & Silvapulle, M. J. Ridge estimation in logistic regression. *Communications in Statistics-Simulation and Computation*, 17(4), 1231-1257, 1988.

[12] Lemonte, A. J., Moreno-Arenas, G., & Castellares, F. Zero-inflated Bell regression models for count data. *Journal of Applied Statistics*, 47(2), 265-286, 2019.

[13] Le Cessie, S., & Van Houwelingen, J. C. Ridge estimators in logistic regression. *Applied Statistics*, 41(1), 191-201, 1992.

[14] Lukman, A. F., Adewuyi, E., Månsson, K., & Kibria, B. M. G. A new estimator for the multicollinear Poisson regression model: simulation and application. *Scientific Reports*, 11(1), 3732, 2021.

[15] Lukman, A. F., Aladeitan, B., Ayinde, K., & Abonazel, M. R. Modified ridge-type for the Poisson regression model: simulation and application. *Journal of Applied Statistics*, 49(8), 2124–2136, 2022.

[16] Månsson, K., & Shukur, G. A Poisson ridge regression estimator. *Economic Modelling*, 28(4), 1475-1481, 2011.

[17] Månsson, K. On ridge estimators for the negative binomial regression model. *Economic Modelling*, 29(2), 178-184, 2012.

[18] Montgomery, D. C., Peck, E. A., & Vining, G. G. Introduction to Linear Regression Analysis. *Wiley*, 2012.

[19] Warton, D. I. Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. *Environmetrics: The Official Journal of the International Environmetrics Society*, 16(3), 275-289, 2005.