



Developing a Semiparametric Zero-Inflated Beta Regression Model Using P-splines: Simulation and Application

Muhammad M. Seliem *, Sayed M. El-Sayed , Mohamed R. Abonazel

*Department of Applied Statistics and Econometrics, Faculty of Graduate Studies
for Statistical Research, Cairo University, Giza, Egypt*

Abstract Analyzing proportional data with excessive zeros and complex relationships presents a significant challenge in various fields. To address this, we propose a developing semiparametric zero-inflated Beta Regression (SPZIBE) regression model using both penalized smoothing (Ps) spline and P-spline (Pb) estimators, referred to as SPZIBE-Ps and SPZIBE-Pb, respectively. This model offers a unique combination of flexibility and interpretability, allowing for the modeling of nonparametric relationships and the identification of factors contributing to zero-inflation. Extensive simulations demonstrate the SPZIBE-Pb model's superior model fit and predictive accuracy compared to existing parametric regression models and semiparametric advanced regression models such as SPZIBE-Ps. The SPZIBE-Pb regression model achieves competitive results on metrics such as the akaike information criterion (AIC), bayesian information criterion (BIC), deviance statistic (DVS) and root mean squared error (RMSE), as confirmed by Monte Carlo simulation studies and real-world applications. The SPZIBE-Pb regression model has broad applications in various fields, including political science, economics, and social sciences. To demonstrate its utility, we applied it to the varieties of democracy (V-Dem) dataset. In conclusion, the SPZIBE-Pb regression model offers a robust and versatile tool for analyzing proportional data with excessive zeros and complex relationships. Its ability to capture both parametric and nonparametric effects, coupled with its interpretability, makes it a valuable asset for researchers across various domains.

Keywords Semiparametric Models, Penalized Spline, Zero-Inflated Models, Proportional Data and V-Dem Dataset

Mathematics Subject Classification: 62E10, 60K10, 60N05

DOI: 10.19139/soic-2310-5070-2220

1. INTRODUCTION

It is well known that generalized linear models can be used to represent a relationship between a response variable and a few predictors, provided that the response variable belongs to the exponential family. However, for the unknown functional relationships between a collection of predictors and a response variable, this approach is not appropriate. Thus, the semiparametric generalized linear regression models are suitable and powerful extensions of generalized linear models that can be used to estimate unknown functional relationships between a collection of predictors and response variables.

Since the beta regression model was first introduced by Ferrari and Cribari-Neto (2004), it has become one of the common distributions that fall under the generalized linear models and is used in cases for modeling rates and proportions. This means the beta regression model is used for modeling a continuous response variable, y , that takes values in the unit interval (0;1). Many authors have modeled data that assume values in the standard unit interval (Bayer and Cribari-Neto, 2017; Algamil and Abonazel, 2022; Abonazel et al., 2022; Abonazel and Taha, 2023). The importance of beta regression is due to its representation of many phenomena in which the data is

*Correspondence to: Muhammad M. Seliem (Email: Seliem.m.m@hims.edu.eg). Department of Business Information System, Higher Institute of Management Sciences, Ministry of Higher Education, New Cairo, Egypt

in the form of proportions and fractions in the open unit interval $(0,1)$. However, proportions data often deviates from a beta distribution because proportions data often includes a nonnegligible number of zeros. Previous studies have found that if the trailing zero is not considered, misleading results are obtained. Thus, the zero-inflated beta (ZIBE) regression model is a suitable and powerful extension of the beta regression model when it suffers from this problem. More recently, several works using the ZIBE regression model have been published. For example (Ospina and Ferrari, 2010, 2012; Baione et al., 2021; Tang et al., 2023; Kaulika and Hajarisman, 2023).

Recently, there has been a noticeable development in semiparametric generalized regression models. For example, Ibacache-Pulgar and Paula (2011) introduced partially linear Student-t models. Yousof and Gad (2017) introduces a novel Bayesian semiparametric logistic regression model, which extends the semiparametric logistic regression model (SLoRM) and improves its estimation process. The study compares Bayesian and non-Bayesian estimation methods for both parametric and semiparametric logistic regression models, applying them to credit scoring data, for more details about prior and posterior distributions, see Seliem et al. (2025). Abonazel et al. (2019) compares the performance of two different smoothing approaches (kernel and spline) in estimating the nonparametric component of a partially linear model. The Speckman estimation method is used for both approaches. The study finds that the spline smoothing approach is generally more efficient than the kernel smoothing approach. El-sayed et al. (2019) proposes a new estimator for the partially linear model based on the B-spline approach. The proposed estimator is compared with the traditional Speckman estimator and Abonazel's estimator. The study finds that the B-spline Speckman estimator is more efficient than the other two estimators in terms of bias and mean squared error. Elgohary et al. (2019) address the challenges of estimating partially linear models when dealing with multicollinearity and the presence of outliers in the data. They propose novel robust-ridge estimators that combine the robustness of least trimmed squares with the bias-reducing properties of ridge regression. These estimators are developed within the framework of spline smoothing for the nonparametric component. Through a Monte Carlo simulation study, the authors demonstrate that their proposed estimators outperform the traditional Speckman-spline estimator in terms of efficiency and reliability, particularly in the presence of multicollinearity and outliers. For more details on handling outliers, see Seliem (2022). Also, Abonazel and Gad (2020) introduced a robust estimation method for semiparametric partially linear models based on partial residuals. Their results showed enhanced performance and robustness compared to conventional methods, especially when dealing with outliers. Abonazel et al. (2021) uses a semiparametric approach to estimate Engel curves and measure food poverty in rural Egypt. The partial linear model is used to estimate the Engel curves, and two estimation methods (double residual estimator and differencing estimator) are compared. The study finds that the semiparametric approach is more flexible than the traditional parametric approach and can provide more accurate estimates of food poverty.

Ibacache-Pulgar et al. (2021) studied semiparametric additive beta regression models and developed the local influence method for these models. Vasconcelos et al. (2022) proposed three semiparametric regression models (additive, additive partial, and semiparametric) based on the odd log-logistic generalized inverse Gaussian distribution. Tapia et al. (2019) studied the semiparametric logistic regression model with influential observations. Logistic regression often performs poorly when dealing with binary data containing an unexpectedly high proportion of zeros. This is due to the model's assumption of a specific outcome distribution that may not accurately represent the real-world data. Araújo et al. (2021) investigates the factors influencing the academic performance of undergraduate business students, measured by the number of failing grades. A semiparametric zero-inflated Negative Binomial (ZINB) regression model was employed to analyze the data, considering various covariates such as work status, dissatisfaction with affirmative action scholarships, and the difficulty of balancing work and study. Li and Lu (2022) introduced a semiparametric zero-inflated Bernoulli regression model to overcome this limitation. Wied (2024) introduces a novel semiparametric distribution regression model with instruments and monotonicity constraints to address the issue of endogeneity. The model provides a flexible and robust approach to estimating the entire conditional distribution of an outcome variable. Fendrich et al. (2024) address the challenge of modeling arsenic contamination in European topsoils, which is often complicated by the presence of censored data. To tackle this issue, they propose a novel coupled generalized additive models for location, scale and shape (GAMLSS) and random forest (RF) model. This innovative approach allows for flexible and robust modeling of the entire distribution of arsenic concentrations, capturing complex relationships with environmental factors. The study's

findings contribute to a better understanding of arsenic pollution and its potential health risks. ZIBE regression models are powerful tools for analyzing proportional data (between 0 and 1). However, their inability to capture nonparametric relationships with the response variable presents a limitation. To address this, we develop a SPZIBE regression model that incorporates penalized splines (both Pb and Ps estimators). This approach combines the interpretability of the ZIBE regression model with the flexibility of nonparametric smoothing, allowing for effective estimation of both parametric and nonparametric effects on proportions. This represents a significant advancement in the field, as existing research has primarily focused on separate applications: zero-inflated regression models for count data and semiparametric regression models for continuous data. Our developed SPZIBE regression model bridges this gap, offering a powerful tool for researchers in various fields like economics, epidemiology, and social sciences, where data often exhibits complex relationships.

This article is organized as follows: In **Section 2** we give a brief sketch of the ZIBE regression model. In **section 3**, we are concerned with the SPZIBE regression model with a penalized spline estimators (both Pb and Ps) for the estimation of parametric and nonparametric components. Simulation studies and results are given in **Section 4** to illustrate the advantages of the proposed estimators when simpler models are inadequate. Application, results, and interpretations to a real dataset are presented in **Section 5**, to explain the flexibility of the introduced the SPZIBE regression model via Pb and Ps estimators. Finally, we offer some conclusions in **Section 6**.

2. Zero-Inflated Beta Regression Model

Since the beta distribution denoted by $\mathbb{B}(\mu, \phi)$ is a member of the exponential family, thus that the fractions, $y_i \in (0, 1)(i = 1, \dots, n)$, are generated independently according to beta distribution, in which the response variable with shape parameters μ and ϕ has a probability mass function (p.m.f.) given by

$$f_{BE}(y_i; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma[(1-\mu)\phi]} y_i^{\mu\phi-1} (1-y_i)^{(1-\mu)\phi-1}, \quad y_i \in (0, 1); i = 1, \dots, n \quad (1)$$

where the $\Gamma(\cdot)$ is the gamma function. According to equation (1), the mean of y_i can be written as

$$\eta_{1i} = g(\mu_i) = x_i^T \beta \quad (2)$$

where $X = (x_1, x_2, \dots, x_p)^T$ denotes the observations on p known covariates, $x_1 = 1$ if the model contains the intercept term, $\beta = (\beta_1, \dots, \beta_p)^T \in \mathbb{R}^p$ is a p -dimensional vector of regression coefficients ($p < n$), $\mu_i = g^{-1}(\eta_i)$ is a function of β , η_i is a linear predictor, and g^{-1} is inverse of $g(\cdot)$ which is a strictly monotonic and twice differentiable link function and . Then, from equation (1) the log-likelihood function based on observed data, $y_i(i = 1, 2, \dots, n)$, apart from constant, can be expressed as:

$$\ell(\beta, \phi) = \sum_{i=1}^n \ell_i(\mu_i, \phi) = \log \Gamma(\phi) - \log \Gamma(\mu_i \phi) - \log \Gamma[(1 - \mu_i) \phi] + (\mu_i \phi - 1) \log y_i + [(1 - \mu_i) \phi - 1] \log(1 - y_i) \quad (3)$$

The MLE is the most used method for the estimation of unknown regression parameters of the beta regression model. Since the equation (3) is nonlinear in β , the solution is obtained using iterative methods. A common such procedure is the iteratively re-weighted least squares (IRLS) method. Let $\beta^{(r+1)}$ be the estimated value of MLE of β with r iterations which may be written as

$$\beta^{(r+1)} = \beta^{(r)} - (I)_{\beta^{(r)}}^{-1} S(\beta) \Big|_{\beta^{(r)}}$$

Subsequently, the estimated coefficients are defined as

$$\hat{\beta} = (X^T \hat{W} X)^{-1} X^T \hat{W} \hat{U} \quad (4)$$

where

$$\hat{U} = \log(\hat{\mu}_i) + (y_i - \hat{\mu}_i) / \sqrt{\text{var}(\hat{\mu}_i)} \quad \text{and} \quad \hat{W} = \text{diag}(\hat{w}_1, \hat{w}_2, \dots, \hat{w}_n)$$

A Zero-one-inflated beta regression model is an alternative way to model fractions data with an excess of zeroes and can be formulated as follows:

$$f_{\Lambda}(y_i; \mu_i, \phi, \pi) = \begin{cases} \pi, & \text{if } y_i = \Lambda \\ (1 - \pi)f_{BE}(y_i; \mu_i, \phi), & \text{if } y_i \in (0, 1) \end{cases}; \Lambda = 0, 1 \quad (5)$$

where the parameter π represents the probability of observing zero ($\Lambda = 0$) or one ($\Lambda = 1$). When $\Lambda = 0$, $f_{\Lambda} = f_{ZIBE}(y_i; \mu, \phi, \pi)$ in equation (5) is known as a ZIBE distribution, suitable for datasets containing excess zeros. Conversely, when $\Lambda = 1$, it's called a one-inflated beta distribution, appropriate for datasets with an abundance of ones. The mean of y and its variance can be written as follows:

$$E(y) = \pi\Lambda + (1 - \pi)\mu$$

and

$$Var(y) = (1 - \pi)\frac{\mu(1 - \mu)}{\phi + 1} + \pi(1 - \pi)(\Lambda - \mu)^2$$

Note that the expected value, $E(y)$, is the weighted average of two components: the mean of a degenerate distribution at Λ (either 0 or 1) and the mean of a beta distribution with parameters μ and ϕ . The weights assigned to these components are π and $1 - \pi$, respectively. Additionally, the conditional expectation of y , given that y lies between 0 and 1, is equal to μ , and its conditional variance is $\frac{\mu(1-\mu)}{1+\phi}$. This article focuses on ZIBE regression model, assuming the response variable follows a mixed continuous-discrete distribution with probability mass at zero. In the context of GLMs, three link functions are commonly employed for ZIBE regression modeling:

$$\begin{aligned} \eta_{1i} &= g_1(\mu_i) = \log\left(\frac{\mu_i}{1 - \mu_i}\right) = x_i^T \beta \\ \eta_{2i} &= g_2(\phi) = \log(\phi) = u_i^T \alpha \\ \eta_{3i} &= g_3(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = z_i^T \omega \end{aligned} \quad (6)$$

where $\beta = (\beta_1, \dots, \beta_p)^T \in \mathbb{R}^p$, $\alpha = (\alpha_1, \dots, \alpha_h)^T \in \mathbb{R}^h$ and $\omega = (\omega_1, \dots, \omega_q)^T \in \mathbb{R}^q$ are vectors of unknown regression coefficients, which are assumed to be functionally independent and $X = (x_1, x_2, \dots, x_p)^T$, $U = (u_1, u_2, \dots, u_h)^T$ and $Z = (z_1, z_2, \dots, z_q)^T$ are observations on p , h , and q known explanatory variables. Also, we assume that the link function g_1 , g_2 and g_3 are strictly monotonic and twice differentiable. There are several possible choices for the link function $g(\cdot)$. For instance, one can use the logit specification. Then, from equations (5) and (6), the likelihood function for the ZIBE regression model, with respect to the parameter vectors β , α and ω , given the observed sample, is given as equation:

$$\begin{aligned} \mathbb{L}(\beta, \alpha, \omega) &= \prod_{i=1}^n f_{\Lambda}(y_i; \mu_i, \phi, \pi) = \mathbb{L}_1(\omega)\mathbb{L}_2(\beta, \alpha); \\ \mathbb{L}_1(\omega) &= \prod_{i=1}^n \pi_i^{\mathbb{I}_{\{0\}}(y_i)} (1 - \pi_i)^{1 - \mathbb{I}_{\{0\}}(y_i)} \text{ and } \mathbb{L}_2(\beta, \alpha) = \prod_{i: y_i \in (0, 1)} f_{BE}(y_i; \mu, \phi) \end{aligned}$$

where $\mathbb{I}_A(y_i)$ being an indicator function that equals 1 if $y_i \in A$ and 0, if $y_i \notin A$ and the log-likelihood function as follows:

$$\ell(\beta, \alpha, \omega) = \ell(\omega) + \ell(\beta, \alpha) = \sum_{i=1}^n \ell_i(\pi_i) + \sum_{i: y_i \in (0, 1)} \ell_i(\mu_i, \phi_i) \quad (7)$$

where

$$\ell_i(\pi_i) = \mathbb{I}_{\{\Lambda\}}(y_i) \log(\pi_i) + [1 - \mathbb{I}_{\{\Lambda\}}(y_i)] \log(1 - \pi_i),$$

$$\ell_i(\mu_i, \phi) = \log \Gamma(\phi_i) - \log \Gamma(\mu_i \phi_i) - \log \Gamma[(1 - \mu_i) \phi_i] + (\mu_i \phi_i - 1) y_i^* + (\phi_i - 2) y_i^{**},$$

where $y_i^* = \log\left(\frac{1-y_i}{y_i}\right)$ and $y_i^{**} = \log(1 - y_i)$ if $y_i \in (0, 1)$, and $y_i^* = 0$ and $y_i^{**} = 0$ otherwise.

Since equation (7) is nonlinear in $\delta = (\beta^T, \alpha^T, \omega^T)^T$, the solution is obtained using iterative methods. A common such procedure is the iteratively re-weighted least squares (IRLS) or expectation-maximization (EM) algorithms. Then, the maximum likelihood estimator (MLE) is noted as

$$\hat{\delta}_{MLE} = (\hat{\beta}^T, \hat{\alpha}^T, \hat{\omega}^T)^T$$

3. Semiparametric Zero-Inflated Beta Regression Model

While the ZIBE regression model is a powerful tool, it can struggle to capture complex, nonparametric relationships between the explanatory variables and the response variable. To address this limitation, we can extend the model to a SPZIBE regression model. This involves introducing a nonparametric function for a specific continuous explanatory variable, denoted by t in equation (5). This nonparametric function allows the model to capture the nonlinear effects of t on the response variable (y) in a data-driven manner. This approach generalizes the ZIBE regression model by providing more flexibility in modeling complex relationships.

Let $Y = (y_1, \dots, y_n)^T$ be independent random variables, where $Y_i \sim ZIBE(\mu, \phi, \pi)$ for $i \in (1, \dots, n)$ and $y = (y_1, \dots, y_n)^T$ are the corresponding observations of Y . Then, we define the SPZIBE regression model structure based on equation (5) by the systematic component expressed as

$$\eta_{4i} = \eta_{1i} + m(t_i) = x_i^T \beta + m(t_i) \tag{8}$$

where $\beta = (\beta_1, \dots, \beta_p)^T \in \mathbb{R}^p$, $X = (x_1, x_2, \dots, x_p)^T$, t is an observable continuous covariate, $m(\cdot)$ represents a smoothing function that flexibly captures the relationship between a continuous explanatory variable and the response variable. The regression structure in equation (5) combined with the systematic component in equation (8) defines the SPZIP because it contains parametric and nonparametric terms after replacing η_{1i} in (5) with η_{4i} in (8). The η_{4i} in equation (8) utilizes a two-part. The first part, η_{1i} , represents a linear predictor related to the mean through a link function, $g(\mu_i)$. The selection of the link function, denoted by $g(\cdot)$, plays a crucial role in generalized linear models (GLMs). Common choices include **logit**: $g(\mu) = \log\left(\frac{1-\mu}{\mu}\right)$, **probit**: $g(\mu) = \Phi^{-1}(\mu)$ and complementary **log-log**: $g(\mu) = \log[-\log(\mu)]$. The second part, $m(\cdot)$, incorporates a smoothing function to capture the nonlinear effects of the continuous explanatory variable and can be estimated by a penalized spline estimators.

P-splines, a type of penalized spline, are piecewise polynomial functions constructed using B-spline basis functions. These basis functions represent the relationship between the nonparametric explanatory variable and the response variable (dependent variable) in a segmented fashion. These basis functions are subject to a penalty term that controls the smoothness of the resulting spline. This penalty term ensures the P-spline avoids excessive wiggleness and better captures the underlying trend in the data. The smooth function $m(\cdot)$ can be approximated by a linear combination of B-spline basis functions. These B-spline functions are defined by a set of knots, which act like control points, influencing the smoothness and shape of the resulting curve. Let (a, b) be an interval and define set of knots $a = \kappa_1, \dots, \kappa_K = b$, the B-spline basis functions of degree zero, denoted by B_s^0 , are defined as follows:

$$B_s^0(t, \kappa) = \begin{cases} 1 & \text{for } \kappa_s \leq t \leq \kappa_{s+1}; \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

For $d > 0$, B-spline basis functions of degree d , denoted by B_s^d , are defined by the recursive formula:

$$B_s^d(t, \kappa) = \frac{t - \kappa_s}{\kappa_{s+d} - \kappa_s} B_s^{d-1}(t, \kappa) + \frac{\kappa_{s+d+1} - t}{\kappa_{s+d+1} - \kappa_{s+1}} B_{s+1}^{d-1}(t, \kappa); s = 1, \dots, K + d + 1 \quad (10)$$

Note that additional $2d + 2$ knots are necessary for constructing the full B-spline basis of degree d . Then the final form of the curve created by a B-spline of degree d is given by

$$m(t, \kappa) = \sum_{s=1}^{K+d+1} \tau_s B_s^d(t, \kappa) \quad (11)$$

The total number of B-spline basis functions used is denoted by $I = K + d + 1$ and τ_i represent the control points of the B-spline curve (Goepp et al., 2018). The SPZIBE regression model can be estimated by maximizing the penalized likelihood function. Then, the semiparametric regression model in equations (6), (7) and (8), are estimated by maximizing the PS spline log-likelihood function as follows:

$$\ell(\delta, m) = \ell(\delta) - \frac{1}{2} \lambda J(m); \quad J(m) = \int_a^b [m^{(2)}(t)]^2 dt \quad (12)$$

where, $\lambda \geq 0$ is smoothing parameter, $J(m)$ is a penalty term, $m^{(2)}$ refers to the second derivatives and $a = t_1 < \dots < t_n = b$. The penalty in equation (12) may be expressed as

$$J(m) = \int_a^b [m^{(2)}(t)]^2 dt = \Upsilon_s^T M_s \Upsilon_s \quad (13)$$

where M_s is a $q_s \times q_s$ positive semidefinite penalty matrix. However, Eilers and Marx (1996) showed that the integration of the square of the k^{th} derivative of $m(t)$ is well by a penalty on finite differences of the coefficients Υ_s with much less effort, namely

$$\int_a^b [m^{(2)}(t)]^2 dt = \Upsilon_s^T P_k \Upsilon_s \quad (14)$$

where D_k of dimension $(n - k) \times n$ and $P_k = D_k^T \times D_k$. More details of the number of knots and the degrees of freedom can be found in Eilers and Marx (1996). Finally, based on equations (12) and (14) the logarithm of the pb likelihood function can be expressed as:

$$\ell(\delta, m) = \ell(\delta) - \Upsilon_s^T P_k \Upsilon_s \quad (15)$$

4. Simulations and Results

A Monte Carlo simulation study assessed the performance of ZIBE and SPZIBE regression model. Model estimation was carried out through penalized log-likelihood optimization using the R statistical software environment with the GAMLSS package. This research focuses on the SPZIBE regression model. Under the simulated scenario, the response variable, y , was generated from a ZIBE distribution characterized by parameters μ , ϕ , and π . The location parameter, μ , was modeled as a linear combination of covariates, $(x_1$ and $x_2)$, and a nonparametric function of a continuous variable, t . A series of simulations were conducted for varying sample sizes ($n = 150, 300,$ and 450) and replicated 1000 times under ZIBE distributions with zero-inflation proportions of 15% and 30%. Covariates data, X and t , were generated according to the specifications outlined in Table 1. Model performance was evaluated using the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), deviance statistic (DVS), and mean squared error (MSE) under diverse conditions detailed in Table 1. The goodness-of-fit of the nonparametric function, $m(t)$, and the linear coefficients, β , were quantified by average estimates (AEs), mean squared error (MSE), and root mean squared error (RMSE). Specifically, the MSEs for

$m(t)$ and β were calculated as:

$$MSE_{\mathcal{L}}(\hat{m}(t)) = \frac{1}{n} \sum_{i=1}^n [\hat{m}(t_i) - m(t_i)]^2; \quad MSE_{\mathcal{L}}(\hat{\beta}) = \frac{1}{N} \sum_{j=1}^N [\hat{\beta}_j - \beta_j]^2$$

where \hat{m} and $\hat{\beta}_r$ are the estimated values of m and β_r , respectively. The study evaluated the performance of several regression models in handling data with excess zeros. The models compared included the SPZIBE with Pb and Ps estimators, the parametric ZIBE model. A comprehensive simulation study assessed model performance across various sample sizes and ZI levels using metrics like mean absolute error (MAE), AIC, BIC, DVS, MSE, and RMSE.

Table 1. The generated variables for simulation and the design of the experiment

Variable	Value
$m(t)$	$0.6 + \sin(3\pi t)$
t	U(-0.5, 0.6)
x_1	U(-0.5, 0.5)
x_2	U(-0.5, 0.5)
(β_1, β_2)	(-0.6, 0.6)
ϕ and π	0.25
Number of explanatory variables (N)	2
Replications (\mathcal{L})	1000
Zero-Inflation Ratios ($ZI\%$)	30% and 50%
Sample size (n)	150, 300, and 450

In our simulation study, the SPZIBE-Pb regression model by automatic knot selection consistently outperformed other models. This superiority was evident in terms of AIC, BIC, DVS, MSE, MAE, and RMSE values, particularly when the percentage of zero-inflation increased. While the parametric ZIBE model also exhibited good performance, the SPZIBE-Pb regression model demonstrated a more robust and accurate estimation, especially in the presence of excess zeroes. Additionally, the SPZIBE-Ps regression model, which incorporates both parametric and nonparametric components, showed competitive results, particularly in terms of MAE and RMSE.

The SPZIBE-Pb regression model consistently outperformed other models, demonstrating superior performance across varying sample sizes. This robustness underscores its applicability to a wide range of datasets. As the sample size increased, the performance of all models improved, but the SPZIBE-Pb regression model consistently maintained its lead. This suggests that its effectiveness is not limited to specific sample sizes. Based on the results presented in Tables (2 to 4) and Figures (1 to 3), the SPZIBE regression model consistently outperformed the other models evaluated including both the parametric ZIBE model and other semiparametric advanced regression models such as SPZIBE-Ps, reinforcing its superiority in handling excess zeros. Its superior performance, consistency across different sample sizes, and ability to balance model complexity and prediction accuracy make it a valuable tool for dealing with data containing excess zeros.

Table 2. Estimated AIC, BIC, DVS, MAE, and RMSE values of all models when $n = 150$.

Model	Estimator	ZI %	AIC	DVS	BIC	MSE	Est.		Parametric part		Nonparametric part	
							$\hat{\beta}_1$	$\hat{\beta}_2$	MAE	RMSE	MAE	RMSE
ZIBE	MLE	15%	117.040	103.040	138.115	1.275	-0.390	0.370	0.141	0.176	–	–
SPZIBE	Pb		-21.807	-50.316	21.107	0.998	-0.610	0.600	0.006	0.007	0.025	0.032
	Ps		0.530	-19.470	30.637	1.001	-0.540	0.620	0.049	0.055	0.226	0.259
ZIBE	MLE	30%	156.235	142.235	177.310	1.285	-0.38	0.38	0.141	0.176	–	–
SPZIBE	Pb		334.938	6.151	78.271	0.993	-0.61	0.61	0.005	0.006	0.021	0.029
	Ps		54.428	34.428	84.534	1.001	-0.53	0.63	0.048	0.055	0.225	0.258

Table 3. Estimated AIC, BIC, DVS, MAE, and RMSE values of all models when $n = 300$.

Model	Estimator	ZI %	AIC	DVS	BIC	MSE	Est.		Parametric part		Nonparametric part	
							$\hat{\beta}_1$	$\hat{\beta}_2$	MAE	RMSE	MAE	RMSE
ZIBE	MLE	15%	229.401	215.401	255.327	1.294	-0.71	0.45	0.149	0.184	–	–
SPZIBE	Pb		-64.366	-95.074	-7.498	0.999	-0.6	0.6	0.014	0.018	0.066	0.081
	Ps		-16.508	-36.508	20.529	1.002	-0.59	0.59	0.052	0.06	0.242	0.275
ZIBE	MLE	30%	307.091	293.091	333.018	1.310	-0.72	0.45	0.149	0.184	–	–
SPZIBE	Pb		52.199	21.288	109.443	0.996	-0.6	0.6	0.014	0.018	0.064	0.081
	Ps		93.043	73.043	130.081	1.005	-0.59	0.58	0.052	0.06	0.242	0.276

Table 4. Estimated AIC, BIC, DVS, MAE, and RMSE values of all models when $n = 450$.

Model	Estimator	ZI %	AIC	DVS	BIC	MSE	Est.		Parametric part		Nonparametric part	
							$\hat{\beta}_1$	$\hat{\beta}_2$	MAE	RMSE	MAE	RMSE
ZIBE	MLE	15%	351.652	337.652	380.417	1.305	-0.6	0.53	0.153	0.189	–	–
SPZIBE	Pb		-104.369	-136.38	-38.598	0.999	-0.59	0.6	0.014	0.017	0.062	0.076
	Ps		-31.714	-51.714	9.378	1.002	-0.53	0.66	0.052	0.06	0.245	0.278
ZIBE	MLE	30%	466.234	452.234	494.999	1.32	-0.61	0.53	0.153	0.189	–	–
SPZIBE	Pb		68.518	36.267	134.783	0.999	-0.59	0.61	0.014	0.017	0.063	0.075
	Ps		131.742	111.741	172.834	1.003	-0.54	0.66	0.052	0.06	0.244	0.277

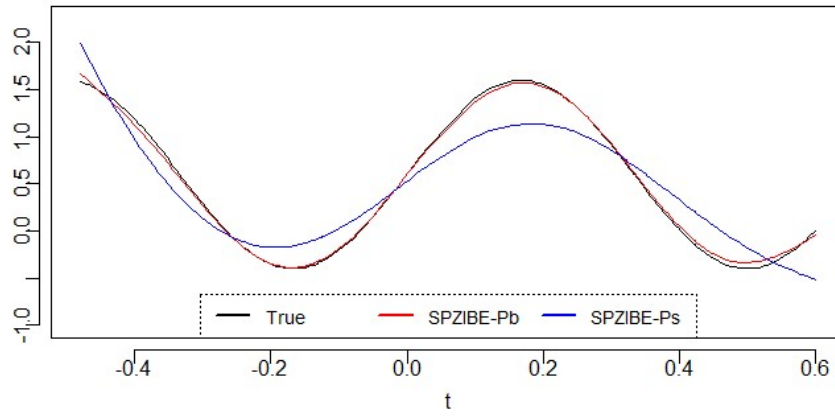


Figure 1. ZI=15%, n=150.

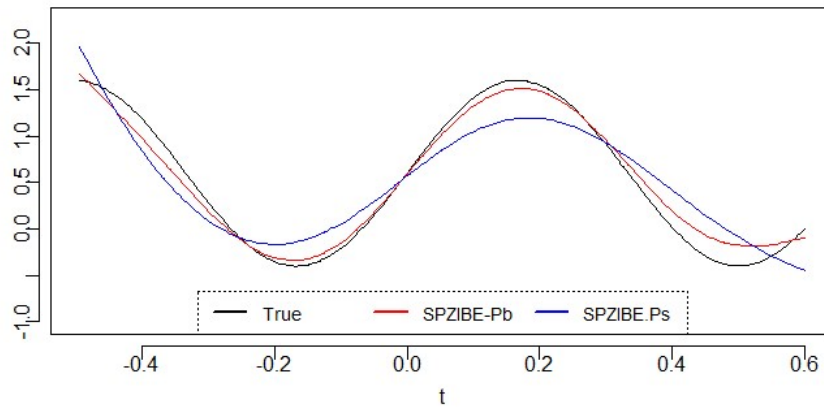


Figure 2. ZI=15%, n=300.

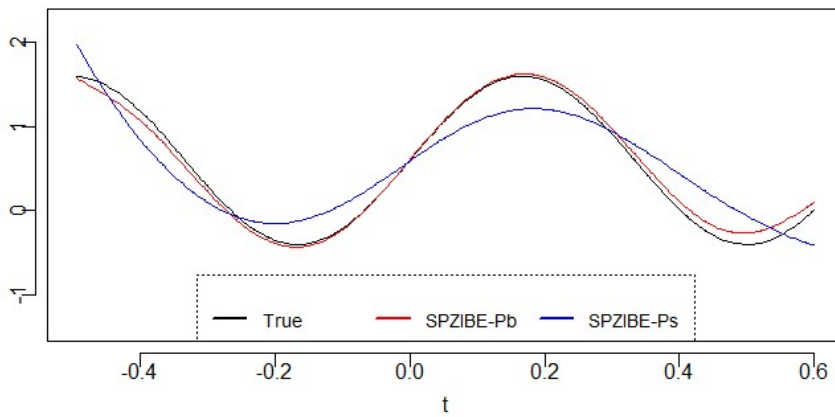


Figure 3. ZI=30%, n=450.

5. Empirical Application

To validate the proposed estimator’s efficacy, this section employs the comprehensive Varieties of Democracy (V-Dem) dataset. There is a group of researchers who used this data, such as [Vaccaro \(2021\)](#), [Treisman \(2023\)](#) and [Ademi and Kimya \(2024\)](#). Our empirical analysis focuses on a selected group of African countries: Mali, Niger, and Burkina Faso to investigate the intricate relationship between political institutions and women’s representation in parliament. Leveraging the V-Dem dataset, we examine the influence of specific political variables on the proportion of women in national legislatures over a historical period spanning from 1950 to 2015. This research offers a novel perspective by applying a semiparametric ZIBE model with P-splines to analyze the relationship between political institutions and women’s representation. By classifying variables, we selected the optimal model, providing a more nuanced understanding of the complex relationships involved. The dataset comprises 134 observations, with one response variable (**Prop-fem**): A measure of the proportion of women in government, typically in parliament or other legislative bodies, and three explanatory variables (x_1 , x_2 and x_3). These explanatory variables are as follows: **Civil Liberties**: A measure of the extent to which individuals can enjoy fundamental freedoms like speech, assembly, and religion without government interference. **Corruption**: An indicator of the perceived level of corruption within a country’s government and public institutions. **Quota**: A binary variable indicating whether or not a country has implemented a quota system to enhance women’s representation in government. By analyzing the interplay between these political variables and the proportion of women in parliament, we aim to shed light on the critical factors influencing women’s political participation in the selected African countries. The descriptive statistics of the data variables in this study are given in Table 5.

Table 5. Descriptive statistics for the variables

Variable names	Description	sample size	Mean	SD
y	Prop-fem		0.047	0.052
x_1	Civil liberties	134	0.560	0.219
x_2	Corruption		0.565	0.195
x_3	Quota	Binary: 1 = quota exists, 0 = quota does not exist).		

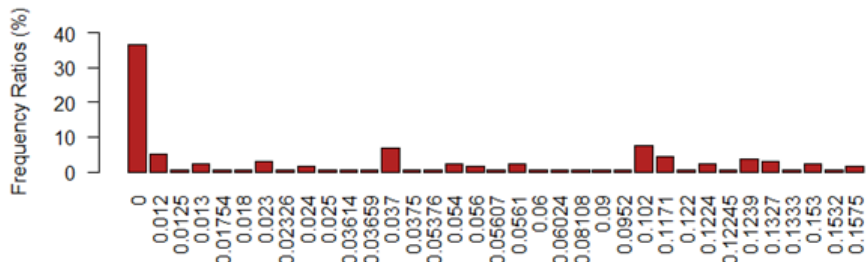


Figure 4. Histograms of the response variable.

The histogram reveals a concerning lack of representation for women in parliaments, with a significant number of countries having no women in their legislative bodies. This underscores the need for policies and initiatives to promote gender equality in politics. As shown in Figure 4, the proportion of zeros in the dependent variable was equal to 36.6%, meaning that 36.6% of the countries or regions represented in the data have no women in their parliaments. This is a significant finding that highlights the underrepresentation of women in political leadership. Figure 5 presents a correlation matrix that illustrates the relationships among three independent variables: x_1 , x_2 and x_3 . The heatmap reveals moderate positive correlations between x_1 and x_2 (0.34).

Furthermore, the correlation between x_1 and x_3 is positive (.34), indicating that quotas might not always promote stronger civil liberties. A weak negative correlation exists between x_2 and x_3 (-0.04), suggesting that the implementation of gender quotas in government has minimal association with corruption levels. Notably, all

Table 6. Checking the relationships between the response and explanatory variables

Models	Estimate	P-value	R-square	Expected Relation
(1) $y \sim x_1$	2.7170	< 0.0001	0.188	Parametric
(2) $y \sim x_2$	0.3581	0.472	0.004	Nonparametric
(3) $y \sim x_3$	0.93274	< 0.0001	0.16	Parametric

correlation coefficients are below 0.8, indicating an absence of multicollinearity among the variables. This is favorable for subsequent regression analyses, as it allows for the accurate assessment of each variable’s unique contribution to the model. Table 6 presents the results of a ZIBE regression model, which analyzes the relationship between y and the predictors x_1 , x_2 , and x_3 . The model reveals a parametric relationship between y and both x_1 and x_3 . However, the relationship with x_2 is nonparametric. This finding is further supported by the low R^2 values for x_2 in Figure 6, suggesting that these variables have limited explanatory power. Based on these results, we recommend including x_1 and x_3 as parametric terms in the final model, while treating x_2 as a nonparametric term.

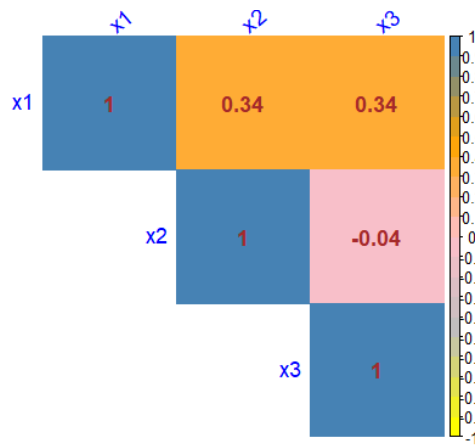


Figure 5. Correlation Matrix.

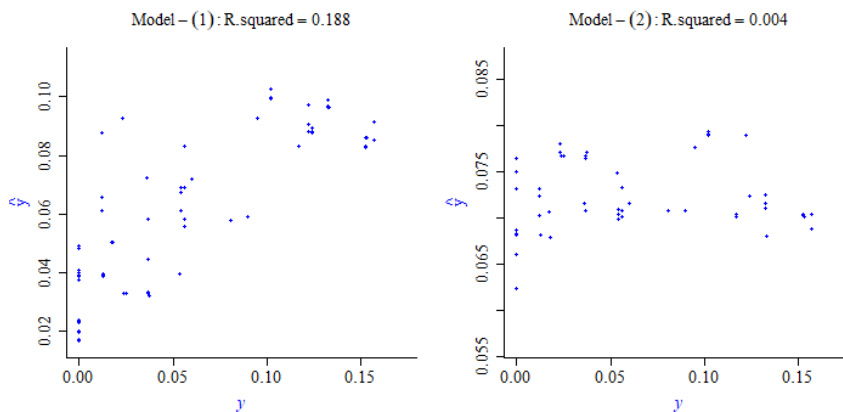


Figure 6. Scatter plots of y VS. \hat{y} .

Table 7 presents the DVS, AIC, BIC, and MSE statistics for the fitted models. The SPZIBE-Pb regression model consistently outperforms the others based on these metrics, indicating a superior fit to the data. Additionally, the R^2 statistics further confirm the SPZIBE-Pb regression model’s efficacy, demonstrating a higher proportion of the data’s variability explained by this model compared to the alternatives.

Table 7. Fitted Regressions Model with Model selection measures

Model	Systematic Components	DVS	AIC	BIC	MSE	R ²
Parametric Model						
ZIBE	$\mu_i = \exp(\beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3)$	-192	-176	-153	1.12	0.44
Proposed-Semiparametric models						
SPZIBE-Pb	$\mu_i = \exp(\beta_0 + \beta_1x_1 + \beta_2x_3) + pb(x_2)$	-457	-404	-328	0.98	0.92
SPZIBE-Ps	$\mu_i = \exp(\beta_0 + \beta_1x_1 + \beta_2x_3) + ps(x_2)$	-363	-324	-267	1.02	0.84

Figure 7 refers to the radar plot that compares the performance of three models (ZIBE, SPZIBE-Pb and SPZIBE-Ps) across three metrics (BIC_{wt} , $RMSE$, AIC_{wt}). Each point on the radar represents a model’s performance for a specific metric. In this plot, values closer to the outer edge indicate better performance. The SPZIBE-Pb regression model emerges as the superior model, positioned closest to the outer edge for all metrics.

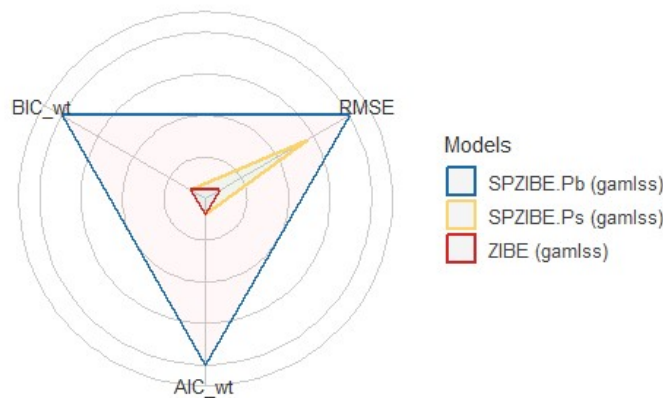


Figure 7. Comparison of Models Performance Indices.

Conversely, the SPZIBE-Ps regression model falls behind, particularly in RMSE where it’s significantly closer to the center. The ZIBE regression model exhibits a balanced performance across the metrics, lacking any major weaknesses. In conclusion, the SPZIBE-Pb regression model demonstrates the best overall performance based on the evaluated criteria. The provided likelihood ratio tests (LRTs) compare the fit of models. In each test, the null model is simpler than the alternative model, with fewer degrees of freedom. Hypotheses based on the LRT results as follows:

Hypothesis 1:

- Null Hypothesis (H_0): The parametric ZIBE regression model is sufficient.
- Alternative Hypothesis (H_1): At least one semiparametric regression model (SPZIBE-Pb or SPZIBE-Ps) is superior.

Hypothesis 2:

- Null Hypothesis (H_0): There is no difference between SPZIBE-Pb and SPZIBE-Ps.
- Alternative Hypothesis (H_1): The SPZIBE-Pb regression model is significantly better.

The likelihood ratio tests (LRTs) in Table 8 strongly support the use of the SPZIBE regression model via Pb estimator. This suggests that Pb estimator captures crucial nonlinear relationships in the data, leading to a more accurate representation of the underlying data patterns compared to other regression models like ZIBE and SPZIBE-Ps. Furthermore, all three LRTs reported in Table 8 show a highly significant rejection of the null hypothesis ($p - value < 0.0001$), indicating that simpler models are insufficient. Tests 1 and 2 demonstrate that both SPZIBE-Pb and SPZIBE-Ps regression models outperform the basic ZIBE regression model. However, Test 3 further highlights the superiority of SPZIBE-Pb regression model, suggesting that the Pb estimator is crucial for accurately capturing the underlying data patterns.

Table 8. LR tests

Tests	Models	Hypothesis	Statistic (w)	P-value
Test 1	ZIBE vs SPZIBE-Pb	Hypothesis 1	265.04	< 0.0001
Test 2	ZIBE vs SPZIBE-Ps		170.46	< 0.0001
Test 3	SPZIBE-Ps VS SPZIBE-Pb	Hypothesis 2	94.581	< 0.0001

Table 9. MLEs, SEs, and P–value for the fitted SPZIBE-Pb regression Model

Variables	Estimate	SE	P-value
Coefficients of μ			
Intercept	-4.6031	0.0767	< 0.0001
x_1	2.154	0.0885	< 0.0001
x_2	0.5515	0.0362	< 0.0001
x_3	0.8352	0.0470	< 0.0001
Coefficients of ϕ			
Intercept	-2.4229	0.4936	< 0.0001
x_1	3.0542	0.5793	< 0.0001
x_2	-2.7139	0.5713	< 0.0001
x_3	-4.7279	0.2482	< 0.0001
Coefficients of π			
Intercept	5.3118	0.2297	< 0.0001
x_2	-11.2625	1.0553	< 0.0001
x_3	-26.9885	20818.47	0.9001

Table 9 presents MLEs, Standard Errors (SEs), and P–values for the fitted SPZIBE-Pb regression model, which accounts for excess zeros. The model includes three components: μ , ϕ , and π , representing the mean, dispersion, and zero-inflation probabilities, respectively. In the μ component, all variables (quotaTRUE, civil liberties, and corruption) exhibit statistically significant positive effects ($p < 0.0001$), with civil liberties having the strongest impact. The ϕ component shows significant effects as well, with quotaTRUE and corruption having negative impacts on dispersion, while civil liberties positively influence dispersion. In the π component, both quotaTRUE and corruption exhibit significant negative effects on the probability of observing zeros. The small standard errors relative to the estimates suggest reliable parameter estimation throughout the model. These findings highlight the complex and nuanced relationships between the predictors and the response variable, emphasizing the importance of employing flexible modeling approaches like the for the fitted SPZIBE-Pb regression model to accurately capture these relationships. Figure 8, a moment bucket plot, visually assesses the distribution of residuals from the fitted SPZIBE-Pb regression model. This plot displays transformed moment skewness against transformed moment excess kurtosis. The shaded area represents the "normal region," where residuals are expected to fall if the model is well-specified and the error terms are normally distributed. In this analysis, most residuals from the SPZIBE-Pb regression model fall within the normal region, indicating a good model fit and suggesting that the error terms are approximately normally distributed. Figure 9(a), known as a detrended transformed Owen’s plot,

is used to assess the distribution of residuals from a statistical model. It plots the detrended transformed Owen’s residuals against the ordered quantile residuals. The shaded area represents confidence intervals, and a horizontal line indicates the expected value under normality. In this specific case, most residuals fall within the confidence intervals, suggesting a good model fit and normally distributed errors. However, a few points near the edges might require further investigation. The Q-statistics plot is a diagnostic tool used to assess the normality of residuals in a centile estimation analysis as shown in Figure 9(b).

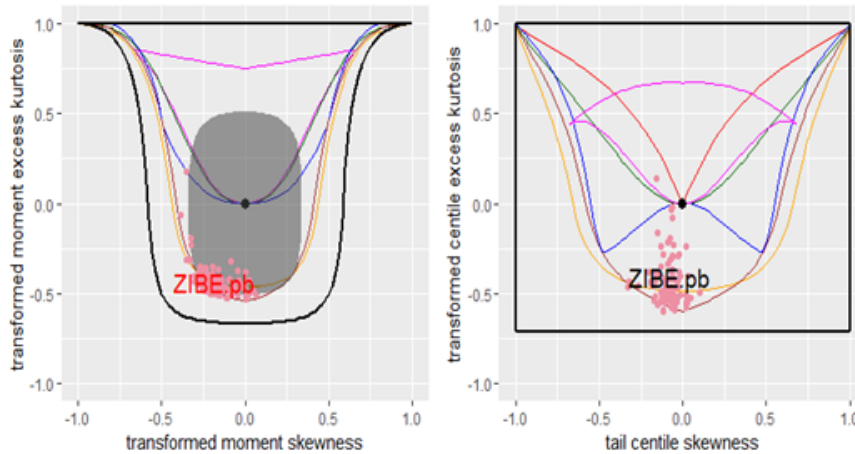


Figure 8. Moment Bucket Plot for Residual Diagnostics.

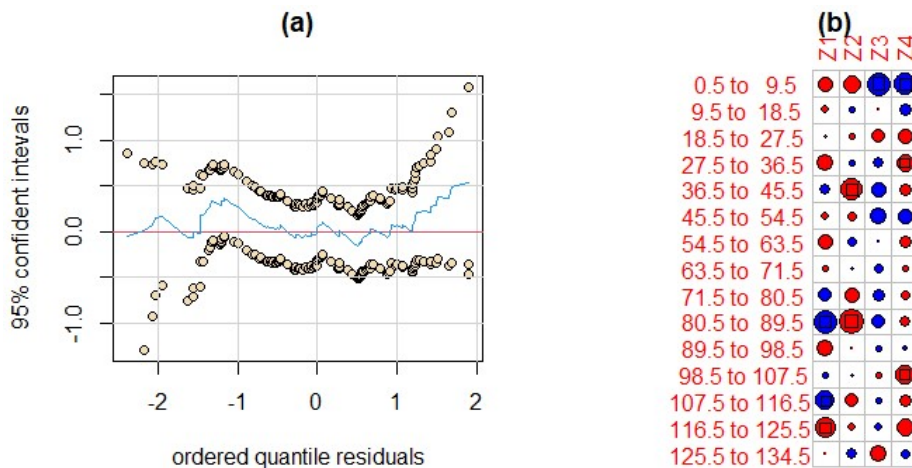


Figure 9. (a) Owen’s plot. (b) Q-statistics plot for Residual Diagnostics.

By visualizing the distribution of residuals across different Z-statistic ranges, the plot helps identify potential deviations from normality. However, to draw definitive conclusions, additional information about the sample size, the meaning of color-coded dots, and the specific context of the analysis is needed. Figure 10 contains four diagnostic plots used to assess residual distribution. Plots (a), (b), and (c), which are normal probability plots, suggest normality of residuals. Plot (d), a residual vs. index plot, shows no clear pattern, indicating random distribution. Overall, the plots indicate normally distributed residuals, suggesting a well-fitted model. Further, the worm plot presented in Panel (a) indicates that there is no evidence of inadequacies in the model since all the residuals fall in the ‘acceptance’ region inside the two elliptic curves. This study has investigated the

complex relationship between political institutions and women’s representation in parliament within selected African countries. By employing a SPZIBE-Pb regression model, we have identified significant nonlinear and linear relationships between civil liberties, corruption, and women’s representation. Our findings demonstrate that increasing civil liberties can lead to a more equitable political landscape. Additionally, higher levels of corruption may be associated with greater female participation in government. While quota systems alone may not directly influence women’s representation, they can create a more conducive environment for female political participation. The semiparametric modeling approach used in this study provides a more flexible and robust analysis than traditional parametric methods. Our results offer valuable insights for policymakers and researchers seeking to promote gender equality and enhance women’s political participation.

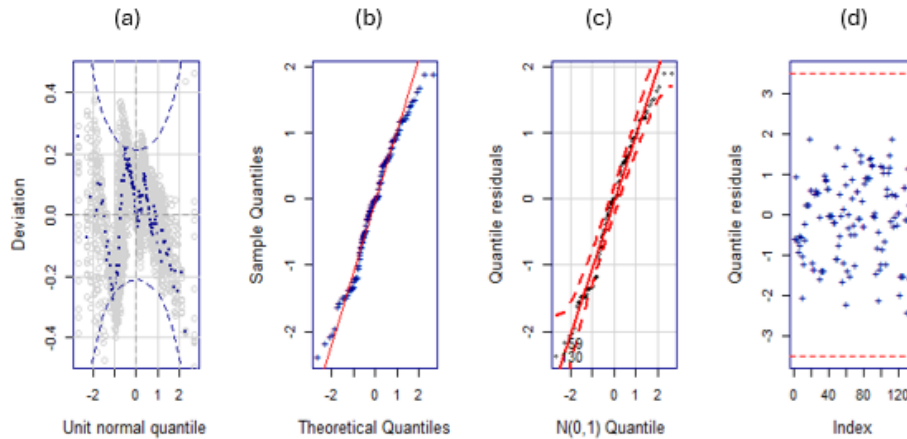


Figure 10. (a) Worm plot. (b-c) Residual Diagnostics. (d) Residuals vs. Order data.

6. Conclusion

This study introduces a SPZIBE regression model that incorporates Pb and Ps estimators to predict a dependent variable influenced by a nonparametric component. The model effectively handles data with zero-inflation levels of 15% and 30%. Comprehensive comparisons using AIC, BIC, deviance, MSE, and RMSE metrics consistently demonstrate the superiority of the proposed model over alternative models, whether parametric models such as ZIBE or semiparametric advanced models such as SPZIBE-Ps regression model. Visualizations further support these findings, showing closer alignment to the true function across various conditions. Simulation results reinforce the model’s robustness and superior performance. The developing SPZIBE regression model with Pb estimator represents a substantial advancement in modeling proportional data characterized by excessive zeros and intricate relationships. By adeptly capturing both linear and nonlinear patterns, this model surpasses existing approaches in terms of model fit and predictive power. Rigorous evaluation using AIC, BIC, DVC, MAE, and RMSE consistently affirmed the superiority of the proposed model. Its flexibility renders it an invaluable tool for researchers across diverse fields confronted with zero-inflated data. Future research should explore extensions to accommodate intricate data structures, such as correlated observations or time-varying covariates, to expand the model’s applicability.

REFERENCES

Abonazel, M., Helmy, N., and Azazy, A. (2019). The performance of speckman estimation for partially linear model using kernel and spline smoothing approaches. *International Journal of Mathematical Archive*, 10(6):10–18.

- Abonazel, M., Shoaeb, F., Abdel Fattah, M., and Abdel-Rahman, S. (2021). Semi-parametric estimation of engel curve and measuring food poverty in rural egypt. *J Appl Probab Stat*, 2021:1–19.
- Abonazel, M. R., Algamal, Z. Y., Awwad, F. A., and Taha, I. M. (2022). A new two-parameter estimator for beta regression model: method, simulation, and application. *Frontiers in Applied Mathematics and Statistics*, 7:780322.
- Abonazel, M. R. and Gad, A. A.-E. (2020). Robust partial residuals estimation in semiparametric partially linear model. *Communications in Statistics-Simulation and Computation*, 49(5):1223–1236.
- Abonazel, M. R. and Taha, I. M. (2023). Beta ridge regression estimators: simulation and application. *Communications in Statistics-Simulation and Computation*, 52(9):4280–4292.
- Ademi, U. and Kimya, F. (2024). Democratic transition and party polarization: A fuzzy regression discontinuity design approach. *Party Politics*, 30(4):736–749.
- Algamal, Z. Y. and Abonazel, M. R. (2022). Developing a liu-type estimator in beta regression model. *Concurrency and Computation: Practice and Experience*, 34(5):e6685.
- Aráujo, E. G., Vasconcelos, J. C., dos Santos, D. P., Ortega, E. M., de Souza, D., and Zanetoni, J. P. (2021). The zero-inflated negative binomial semiparametric regression model: application to number of failing grades data. *Annals of Data Science*, pages 1–16.
- Baione, F., Biancalana, D., and De Angelis, P. (2021). An application of zero-one inflated beta regression models for predicting health insurance reimbursement. In *Mathematical and Statistical Methods for Actuarial Sciences and Finance: eMAF2020*, pages 71–77. Springer.
- Bayer, F. M. and Cribari-Neto, F. (2017). Model selection criteria in beta regression with varying dispersion. *Communications in Statistics-Simulation and Computation*, 46(1):729–746.
- Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical science*, 11(2):89–121.
- El-sayed, S. M., Abonazel, M. R., and Seliem, M. M. (2019). B-spline speckman estimator of partially linear model. *IJSSAM*, 4:53.
- Elgohary, M. M., Abonazel, M. R., Helmy, N. M., and Azazy, A. R. (2019). New robust-ridge estimators for partially linear model. *International Journal of Applied Mathematical Research*, 8(2):46–52.
- Fendrich, A. N., Van Eynde, E., Stasinopoulos, D. M., Rigby, R. A., Mezquita, F. Y., and Panagos, P. (2024). Modeling arsenic in european topsoils with a coupled semiparametric (gamlss-rf) model for censored data. *Environment International*, 185:108544.
- Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of applied statistics*, 31(7):799–815.
- Goepf, V., Bouaziz, O., and Nuel, G. (2018). Spline regression with automatic knot selection. *arXiv preprint arXiv:1808.01770*.
- Ibacache-Pulgar, G., Figueroa-Zuniga, J., and Marchant, C. (2021). Semiparametric additive beta regression models: Inference and local influence diagnostics. *REVSTAT-Statistical Journal*, 19(2):255–274.
- Ibacache-Pulgar, G. and Paula, G. A. (2011). Local influence for student-t partially linear models. *Computational Statistics & Data Analysis*, 55(3):1462–1478.
- Kaulika, L. and Hajarisman, N. (2023). Implementasi zero inflated beta regression model pada proporsi kematian ibu di kota bandung tahun 2020. In *Bandung Conference Series: Statistics*, volume 3, pages 226–235.
- Li, C.-S. and Lu, M. (2022). Semiparametric zero-inflated bernoulli regression with applications. *Journal of Applied Statistics*, 49(11):2845–2869.
- Ospina, R. and Ferrari, S. L. (2010). Inflated beta distributions. *Statistical papers*, 51:111–126.
- Ospina, R. and Ferrari, S. L. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, 56(6):1609–1623.
- Seliem, M. M. (2022). Handling outlier data as missing values by imputation methods: application of machine learning algorithms. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 13(1):273–286.

- Seliem, M. M., Kamel, A. R., Taha, I. M., and El-Nasr, M. M. A. (2025). A note on prior selection in bayesian estimation. *Statistics, Optimization & Information Computing*, 13(1):795–806.
- Tang, B., Frye, H. A., Gelfand, A. E., and Silander, J. A. (2023). Zero-inflated beta distribution regression modeling. *Journal of Agricultural, Biological and Environmental Statistics*, 28(1):117–137.
- Tapia, A., Leiva, V., Diaz, M. d. P., and Giampaoli, V. (2019). Influence diagnostics in mixed effects logistic regression models. *Test*, 28(3):920–942.
- Treisman, D. (2023). How great is the current danger to democracy? assessing the risk with historical data. *Comparative Political Studies*, 56(12):1924–1952.
- Vaccaro, A. (2021). Comparing measures of democracy: statistical properties, convergence, and interchangeability. *European Political Science*, 20(4):666–684.
- Vasconcelos, J. C. S., Cordeiro, G. M., and Ortega, E. M. M. (2022). The semiparametric regression model for bimodal data with different penalized smoothers applied to climatology, ethanol and air quality data. *Journal of Applied Statistics*, 49(1):248–267.
- Wied, D. (2024). Semiparametric distribution regression with instruments and monotonicity. *Labour Economics*, 90:102565.
- Yousof, H. M. and Gad, A. M. (2017). Bayesian semi-parametric logistic regression model with application to credit scoring data. *Journal of Data Science*, 15(1):25–39.