Wiley
Journal of Mathematics
Volume 2024, Article ID 4081833, 18 pages
https://doi.org/10.1155/2024/4081833

WILEY

Research Article

A New Zero–Inflated Regression Model with Applications to Australian Health Survey and Biochemistry Graduate Students Data

Caner Tanış,[1] **Mahmoud M. Mansour**,[2,3] **Enayat M. Abd Elrazik**,[2,3] **Christophe Chesneau**,[4] **Hazem Al-Mofleh**,[5,6] **and Ahmed Z. Afify**[3]

[1]*Department of Statistics, Çankırı Karatekin University, Çankırı 18100, Türkiye*
[2]*Department of Management Information Systems, Taibah University, Yanbu 46422, Saudi Arabia*
[3]*Department of Statistics, Mathematics, and Insurance, Benha University, Benha 13511, Egypt*
[4]*Department of Mathematics, LMNO, CNRS-Université de Caen-Normandie, Campus II, Science 3, Caen 14032, France*
[5]*Department of Mathematical Sciences, Ball State University, Muncie, Indiana 47306, USA*
[6]*Department of Mathematics, Tafila Technical University, Tafila 66110, Jordan*

Correspondence should be addressed to Ahmed Z. Afify; ahmed.afify@fcom.bu.edu.eg

Received 26 September 2023; Revised 25 April 2024; Accepted 22 July 2024

Academic Editor: Ding-Xuan Zhou

In this study, we propose a new zero-inflated regression model as an alternative to zero-inflated regression models, such as the zero-inflated Poisson, zero-inflated negative binomial, zero-inflated hurdle-Poisson, and zero-inflated hurdle negative binomial models. In this regard, we take benefit of the flexibility of the Poisson–Bilal distribution and some of its notable properties. More concretely, it is employed as the baseline distribution to generate a new regression model called the zero-inflated Poisson-Bilal regression model. It is designed to be a good alternative for modeling overdispersed data quite effectively. This aspect is emphasized using two real-world data sets from the medicine and education fields. Furthermore, these data sets are analyzed to compare the goodness-of-fit of the suggested zero-inflated regression model with some of its direct competitors.

1. Introduction

Recently, many studies have aimed to generate novel count regression models as alternatives to current ones. The reason is that many existing count regression models are not adequate to analyze some kinds of data. In particular, the Poisson regression model is a widely used count regression model in many studies [1, 2], but it has some limits when dealing with overdispersed data. The phenomenon of overdispersion is commonly observed in count data and poses significant challenges in statistical modeling. Using the Poisson distribution as the assumed distribution for the dependent variable leads to an underestimation of standard errors and weakens the significance of regression parameters. In such cases, a quasi-Poisson regression model can be employed to overcome the challenge posed by overdispersion. Moreover, researchers have proposed alternative models to deal with the overdispersion, including Poisson-inverse Gaussian, Poisson-lognormal, and negative binomial (NB) regression models [1, 3].

In the last decades, there have been many papers associated with suggesting a regression model in the literature. Some of these studies are listed below. Shanker and Fesshaye [4] introduced the Poisson–Sujatha distribution. Shanker [5] suggested the Poisson–Aradhana distribution. Shanker et al. [6] proposed the Poisson–Akash distribution. Altun [7] created the Poisson-modified Lindley distribution.

On the other hand, count data may exhibit an additional phenomenon referred to as zero-inflation, wherein the occurrence of zero outcomes exceeds the anticipated frequency as dictated by a Poisson distribution. The zero-inflated Poisson (ZIP) and zero-inflated NB (ZINB) regression models, known for their enhanced sensitivity compared to Poisson regression model, are employed in analyzing zero-inflated data sets [2, 8].

Recently, some of the proposed zero-inflated regression models are listed below.

Altun [1] introduced the zero-inflated Poisson–Lindley regression model. Lemonte et al. [9] suggested the zero-inflated Bell regression model. Tanış et al. [2] proposed the zero-inflated discrete Lindley regression model. All of them have qualities and can be revealed to be the best for certain data. However, there is room for improvement or originality with the consideration of simple and flexible intermediary lifetime distributions. Tajuddin et al. [10] studied negative binomial Lindley distribution properties and applications to overdispersed distribution. Tajuddin et al. [11] proposed a new zero–one-inflated Poisson–Lindley Distribution.

This study is devoted to a novel zero-inflated regression model, namely, the zero-inflated Poisson–Bilal regression model, as an alternative to the ZIP and ZINB regression models, with the aim of successfully modeling overdispersed data. As indicated by its name, we employ the discrete Poisson–Bilal (PB) distribution, first introduced by Altun [12], to formulate a zero-inflated count regression model. Thanks to the simplicity and flexibility of the PB distribution, we can effectively analyze a wide variety of overdispersed data. The usefulness of the proposed regression model is presented through two real-world data examples.

The paper is organized as follows: in Section 2, some properties of the discrete PL distribution introduced by Altun [12] are given. Section 3 provides the PB regression model as described by Altun [12]. Section 4 proposes the novel zero-inflated regression model. In Section 5, an extensive Monte Carlo simulation study is performed. Furthermore, Section 6 presents two real data applications for this model and competitor ones. Finally, the concluding remarks are given in Section 7.

2. Discrete PB Distribution

The PB distribution was suggested by Abd-Elrahman, [13]. Some backgrounds are necessary to present it in detail. To begin, the cumulative probability function (CDF) and probability density function (PDF) of the Bilal (B) distribution are given by

$$F(x;\alpha) = 1 - \exp(-2\alpha x)(3 - 2\exp(-\alpha x)), \tag{1}$$

$$f(x;\alpha) = 6\alpha\exp(-2\alpha x)(1 - \exp(-\alpha x)), \quad x > 0 \text{ and } \alpha > 0, \tag{2}$$

respectively. Altun [12] compounded the Poisson distribution with the B distribution to generate a new mixed-Poisson distribution. The probability mass function (PMF), CDF, and hazard rate function (HRF) of the compounded PB distribution are thus obtained as

$$p(x;\alpha) = 6\alpha\left(\frac{1}{(2\alpha+1)^{x+1}} - \frac{1}{(3\alpha+1)^{x+1}}\right), \tag{3}$$

$$\Pr(x;\alpha) = 1 + 2(3\alpha+1)^{-x-1} - 3(2\alpha+1)^{-x-1}, \tag{4}$$

and

$$h(x;\alpha) = \frac{6\alpha\left((2\alpha+1)^{-x-1} - (3\alpha+1)^{-x-1}\right)}{3(2\alpha+1)^{-x-1} - 2(3\alpha+1)^{-x-1}}, \quad x = 0, 1, 2, \ldots \text{ and } \alpha > 0, \tag{5}$$

respectively. Altun [12] emphasized that the PB distribution can be useful in modeling extremely right-skewed and zero-inflated count data sets. Figures 1 and 2 provide some shapes of the PMFs and HRFs for the PB distribution, respectively.

2.1. Some Characteristics of the PB Distribution. In this subsection, we examine some distributional properties of the PB distribution. The first moment, variance, and index of dispersion for a random variable X with this distribution are

$$\mu = E(X) = \frac{5}{6\alpha}, \tag{6}$$

$$\sigma^2 = \mathrm{Var}(X) = \frac{30\alpha + 13}{36\alpha^2}, \tag{7}$$

and

$$\mathrm{ID}(X) = \frac{\mathrm{Var}(X)}{E(X)} = 1 + \frac{13}{30\alpha}, \tag{8}$$

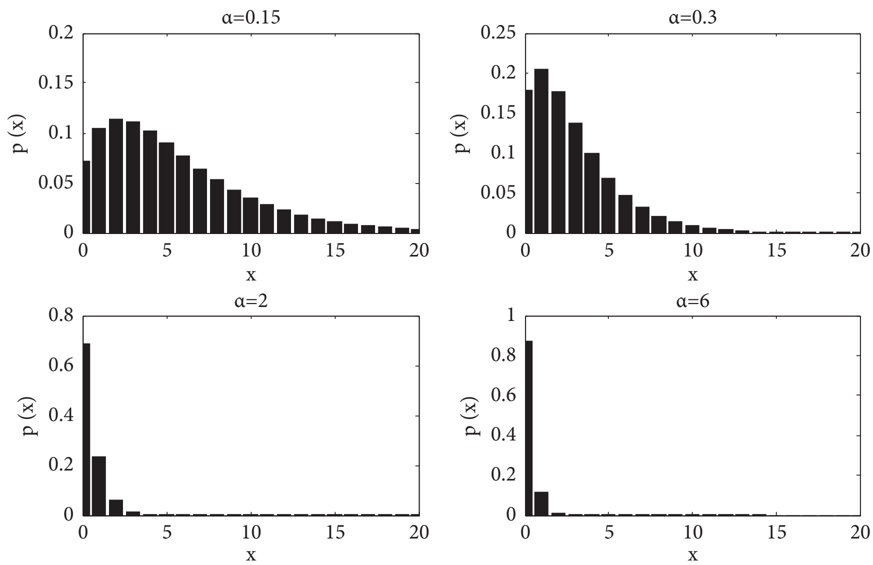

Figure 1: Some PMFs of the PB distribution for selected α values.

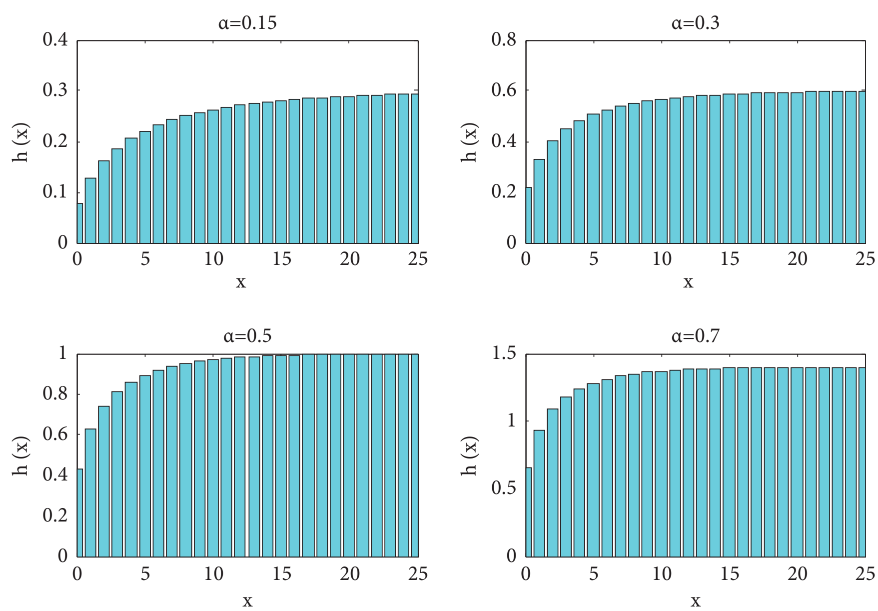

Figure 2: Some HRFs of the PB distribution for selected α values.

respectively. If $\mathrm{ID} \geq 1$ in other words, $\mathrm{Var}(X) > E(X)$ since $E(X) > 0$, the distribution has the overdispersion property [12, 14]. This is clearly the case for the PB distribution. Thus, this property motivates us to generate a new zero-inflated model using the PB distribution as described in the following section.

3. PB Regression Model

The PB regression model was introduced in [12]. More precisely, Altun [12] emphasized that the PB regression distribution provides usefulness in modeling the overdispersed data sets. In this regard, the PB distribution is a good competitor to the Poisson and NB distributions.

Using the reparametrization, $\alpha = 5/(6\mu)$ (inverse of the mean of the $\mathrm{PB}(\alpha)$ distribution in (6)), the PMF of the PB distribution is indicated as

$$p(y;\mu) = \frac{5}{\mu}\left(\left(\frac{5}{3\mu}+1\right)^{-y-1} - \left(\frac{5}{2\mu}+1\right)^{-y-1}\right), \tag{9}$$

where $\mu > 0$ and the covariates are linked to the response variable through log-link function, $\log(\mu_i) = x_i^T\beta$, $\beta = (\beta_1, \beta_2, \ldots, \beta_k)^T \in \mathbb{R}^k$, $x_i^T = (x_{i1}, x_{i2}, \ldots, x_{ik})$.

3.1. Parameter Estimation on the PB Regression Model. Here, we provide the parameter estimation of the PB regression model by using the maximum likelihood method. Let $y_i\,(i = 1, 2, \ldots, n)$ be the observed value of the dependent variable associated with the PB regression model. The corresponding log-likelihood function is given by

$$\ell(\beta) = n\log(5) - \sum_{i=1}^{n} x_i^{T}\beta + \sum_{i=1}^{n}\log\left[\left(\frac{5}{3\exp(x_i^{T}\beta)} + 1\right)^{-y_i-1} - \left(\frac{5}{2\exp(x_i^{T}\beta)} + 1\right)^{-y_i-1}\right]. \tag{10}$$

The MLE of β, say $\widehat{\beta}$, can be described as follows:

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmax}}(\ell(\beta)), \tag{11}$$

and it can be obtained by the optim function in *R* software with some algorithm namely, Nelder-Mead, BFGS, and CG.

4. Zero-Inflated Poisson-Bilal (ZIPB) Regression Model

In this section, we suggest a new count regression model called the zero-inflated Poisson–Bilal (ZIPB) regression model as an alternative to some well-known ones, such as the Poisson, NB, PB, ZIP, and ZINB regression models.

Let U be a Bernoulli random variable with $1-\pi$ success probability, X be a random variable with the $\mathrm{PB}(\alpha)$ distribution, and Y be a random variable defined as follows:

$$Y = \begin{cases} 0, & U = 0, \\ X, & U = 1. \end{cases} \tag{12}$$

Then, the zero-inflated $\mathrm{PB}(\alpha)$ distribution is described by the PMF of the random variable Y in (12) as follows:

$$P(Y = y) = \begin{cases} \pi + (1-\pi)6\alpha\left(\dfrac{1}{(2\alpha+1)} - \dfrac{1}{(3\alpha+1)}\right), & y = 0, \\ (1-\pi)6\alpha\left(\dfrac{1}{(2\alpha+1)^{y+1}} - \dfrac{1}{(3\alpha+1)^{y+1}}\right), & y > 0, \end{cases} \tag{13}$$

where $0 < \pi < 1$ and it is denoted $\mathrm{ZIPB}(\alpha, \pi)$. Also, it is noticed that the random variable given in (12) can be represented by $Y = UX$, where U and X are independent random variables having the $\mathrm{Bernoulli}(1-\pi)$ and $\mathrm{PB}(\alpha)$ distribution, respectively. In this regard, we provide the mean and variance of Y as

$$E(Y) = (1-\pi)\mu, \tag{14}$$

and

$$\mathrm{Var}(Y) = (1-\pi)\left(\sigma^2 + \mu^2\pi\right), \tag{15}$$

respectively, where μ and σ^2 are mean and variance associated with the $\mathrm{PB}(\alpha)$ distribution presented in (6) and (7).

Using the reparametrization, $\alpha = 5/(6\mu)$ in (9), the PMF of the ZIPB distribution is

$$P(Y = y) = \begin{cases} \pi + (1-\pi)\dfrac{25}{(3\mu+5)(2\mu+5)}, & y = 0, \\ (1-\pi)\dfrac{5}{\mu}\left(\left(\dfrac{5}{3\mu} + 1\right)^{-y-1} - \left(\dfrac{5}{2\mu} + 1\right)^{-y-1}\right), & y > 0. \end{cases} \tag{16}$$

The ZIPB count regression model is given by following equations:

$$\log(\mu_i) = x_i^{T}\beta, \tag{17}$$

and

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = z_i^{T}\gamma, \tag{18}$$

where $\mu > 0, \beta = (\beta_1, \beta_2, \ldots, \beta_k)^T \in \mathbb{R}^k, \gamma = (\gamma_1, \gamma_2, \ldots, \gamma_m)^T \in \mathbb{R}^m$.

Also, $x_i^T = (x_{i1}, x_{i2}, \ldots, x_{ik})$ and $z_i^T = (z_{i1}, z_{i2}, \ldots z_{im})$. In the ZIPB count regression model, the mean associated with the $PB(\alpha)$ distribution μ_i is linked to covariates x_i by the log link function and π_i is linked to z_i by the logit link function.

4.1. Parameter Estimation of the ZIPB Regression Model. In this part, we provide the parameter estimation of the ZIPB regression model by using the maximum likelihood method. The corresponding log-likelihood function is given by

$$\ell(\psi) \propto \sum_{i\in\{j:\ y_j=0, j=1,2,\ldots n\}} \log\left[\frac{\exp\left(z_i^T\gamma\right)}{1+\exp\left(z_i^T\gamma\right)} + \frac{25}{\left[1+\exp\left(z_i^T\gamma\right)\right]\left(3\exp\left(x_i^T\beta\right)+5\right)\left(2\exp\left(\left(x_i^T\beta\right)\right)+5\right)}\right]$$
$$+ \sum_{i\in\{j:\ y_j>0, j=1,2,\ldots n\}} \log\left[\frac{5}{\left(1+\exp\left(z_i^T\gamma\right)\right)\exp\left(\left(x_i^T\beta\right)\right)}\left(\left(\frac{5}{3\exp\left(\left(x_i^T\beta\right)\right)}+1\right)^{-y_i-1} - \left(\frac{5}{2\exp\left(\left(x_i^T\beta\right)\right)}+1\right)^{-y_i-1}\right)\right], \tag{19}$$

where $\psi = (\beta, \gamma)$. The MLE of ψ, say $\widehat{\psi}$, can be described as follows:

$$\widehat{\psi} = \underset{\psi}{\operatorname{argmax}}(\ell(\psi)), \tag{20}$$

and it can be obtained by optim function in R software with different algorithms such as Nelder-Mead, BFGS, and CG.

5. Simulation Study for ZIPB Count Regression Model

To assess the performance and explore the behavior of the MLEs, a Monte Carlo simulation is performed in this section, with respect to MLEs average of biases and MSEs. All simulations were repeated 20000 trials for each sample size = $\{100, 200, 300, 400, 500, 600, 700, 800, 900, 1000\}$. The following true parameters settings are considered:

$\theta = 0.5$, $\beta = \{\beta_1 = (0.6, 0.9, 0.5, 3.0)^T, \beta_2 = (0.1, 0.1, 0.1, 0.1)^T\}$, and $\gamma = \{\gamma_1 = (0.5, 0.8, 0.4, 2.5)^T, \gamma_2 = (0.2, 0.2, 0.2, 0.2)^T\}$. The covariates $x_i^T = (1, x_{i1}, x_{i2}, x_{i3})$ and $z_i^T = (1, z_{i1}, z_{i2}, z_{i3})$ are generated from a multivariate normal distribution with mean zero and the following correlation matrices:

$$\rho_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$
$$\rho_2 = \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}. \tag{21}$$

Thus, the variances are fixed at 1. Thus, it can be observed the multicollinearity effect on the ZIPB regression analysis. These simulations are performed based on the following algorithm:

A1. For a fixed sample size, n, the covariates are generated from a multivariate normal distribution with vector mean 0 and one of the presented correlation matrices (ρ_1 or ρ_2). The R programming statistical language function mvrnorm within the package MASS is used to generate those covariates.

A2. Compute $\mu_i = e^{x_i^T\beta}$ and $\pi_i = e^{z_i^T\gamma}/1 + e^{z_i^T\gamma}$, $i = 1, 2, \ldots, n$ for the true values of the parameters under consideration.

A3. For $i = 1, 2, \ldots, n$, generate the dependent variable Y_i from the ZIPB(μ_i, π_i) distribution using the accepting rejection sampling.

A4. The numerical methods are used to maximize the log-likelihood given in (19).

From Figures 3, 4, 5, and 6, it can be concluded that the MSEs decrease as n increases and the biases decrease to zero as n increases, for each parameter combination and multicollinearity level.

6. Practical Data Analysis

In this section, we provide two practical data examples to illustrate the superiority of the proposed regression model over some well-known ones.

6.1. Australian Health Survey Data. The Australian Health Survey (AHS) data set comprises 5190 observations for each covariate obtained from the faraway package in R software. The AHS data set is denoted as dvisits. It is anonymized, and it is open-access available in [15]. In this data set, we consider the response variable to be the number of appointments in the past two weeks with a doctor or specialist. We provide the considered response variable and covariates of the count regression model in Table 1.

Figure 7 illustrates the histogram of the number of appointments in the AHS data set.

We analyze this data set with the Poisson, NB, PB, ZIP, ZINB, zero-hurdle-Poisson, zero-hurdle-NB, and ZIPB regression models. The MLEs and corresponding standard errors (SEs) of the fitted regression models are given in Tables 2, 3, 4, and 5. Table 6 provides the selection criteria statistics.

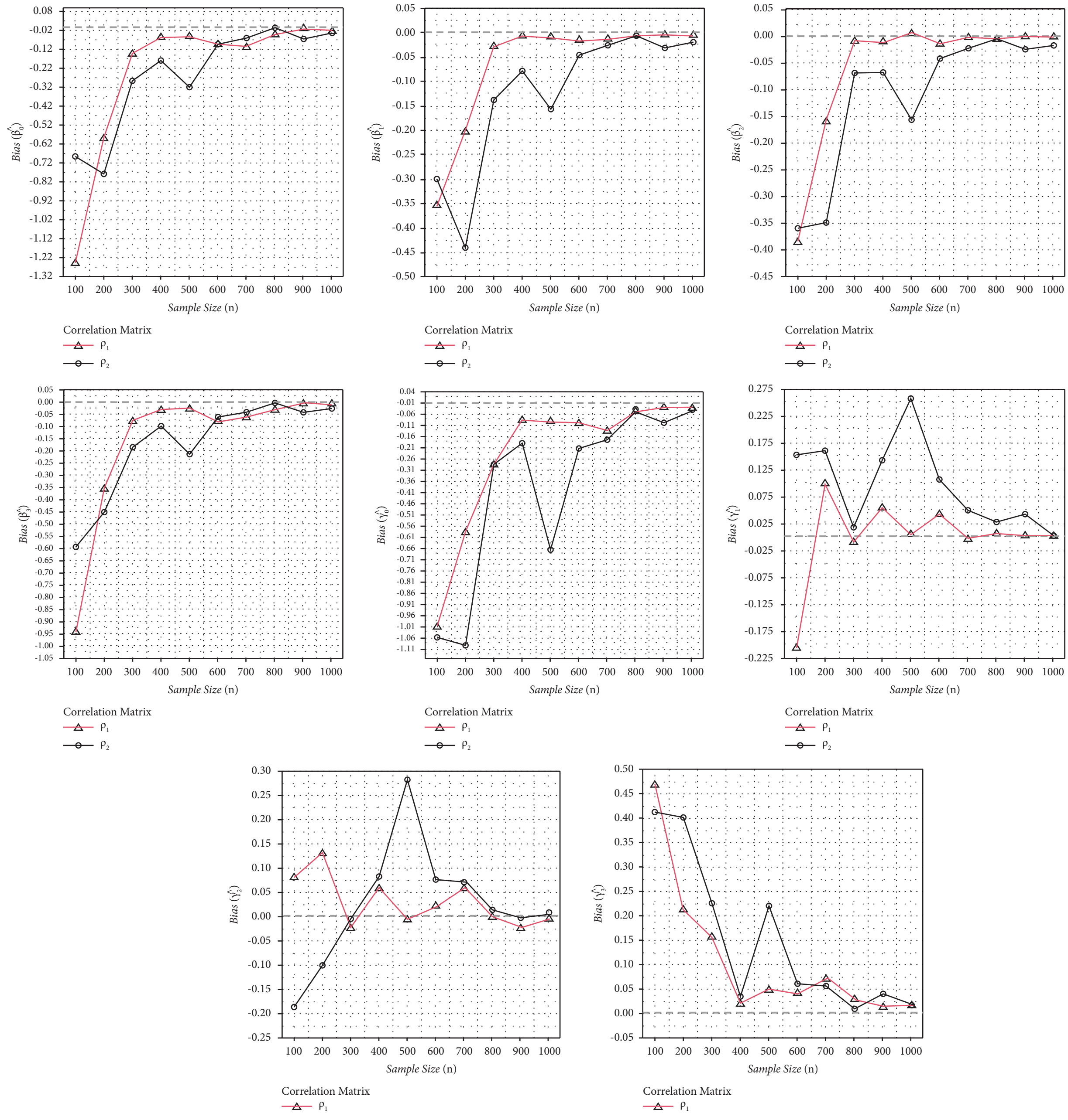

Figure 3: Continued.

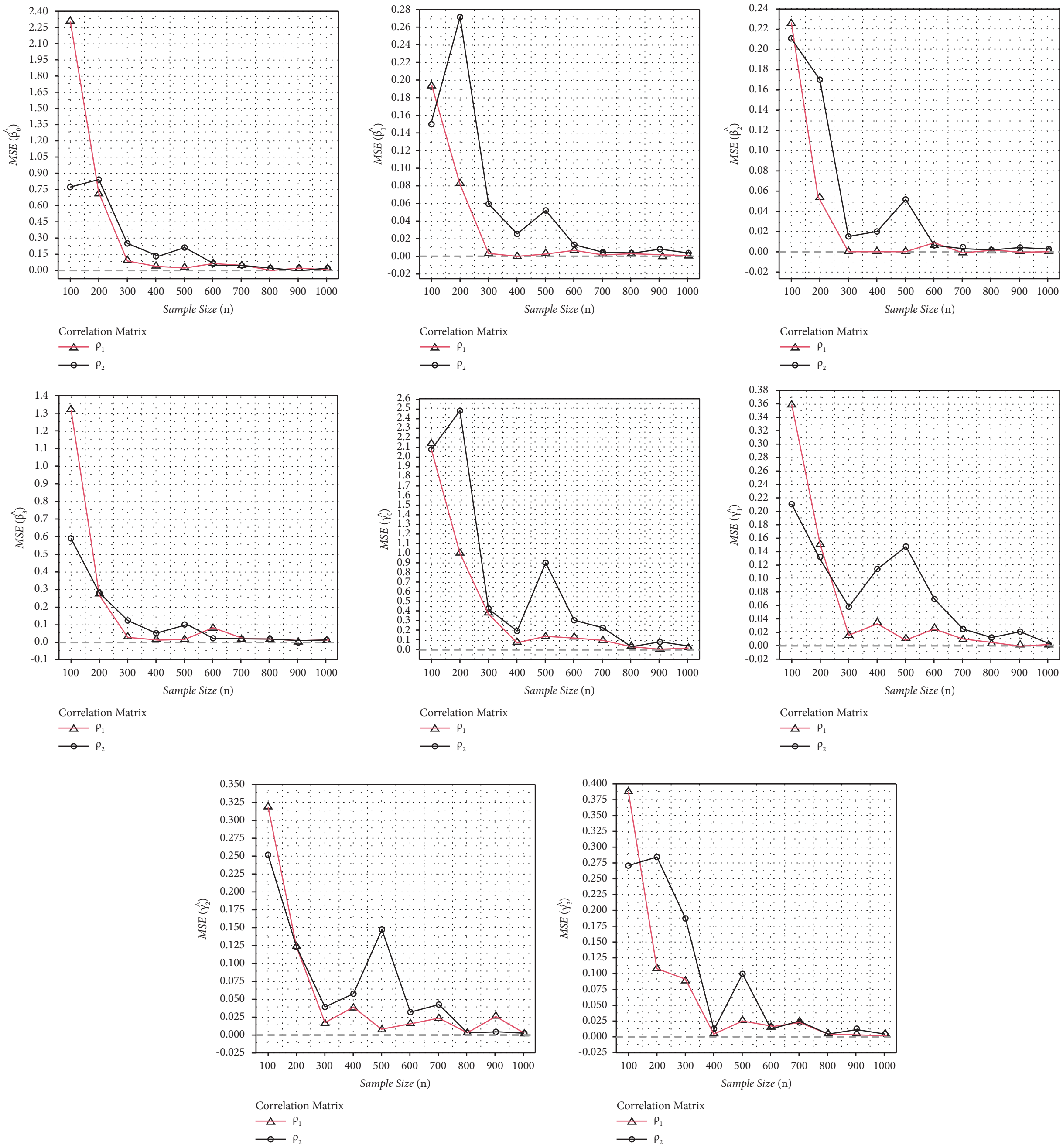

Figure 3: The biases and MSEs of the ZIPB count regression coefficients for the true parameters $\beta_1 = (0.6, 0.9, 0.5, 3.0)$ and $\gamma_1 = (0.5, 0.8, 0.4, 2.5)$ and the considered different correlation matrices.

According to the results of the data analysis, the coefficient β_0 is statistically significant in all the regression models analyzed except for the zero-hurdle-NB regression model ($p < 0.05$). The coefficient β_1 is statistically significant in all the regression models except for the ZIP, ZIPB, and zero-hurdle-NB regression models. In other words, "sex" has a statistically significant effect on "doctorco" according to the Poisson, NB, PB, ZINB, and zero-hurdle Poisson regression models. "age" (β_2) has a statistically significant effect on "doctorco" according to the Poisson, NB, PB, ZINB, and zero-hurdle Poisson regression models. "income" (β_3) has a statistically significant effect on "doctorco" according to only the zero-hurdle NB regression model. "illness" (β_4) is statistically significant in all the models except for the ZIP, zero-hurdle NB, and ZIPB regression models. "actdays" (β_5) is statistically significant in all the examined regression models. "hscore" (β_6) has a statistically significant effect on "doctorco" according to the Poisson, NB, PB, and ZINB regression models ($p < 0.05$).

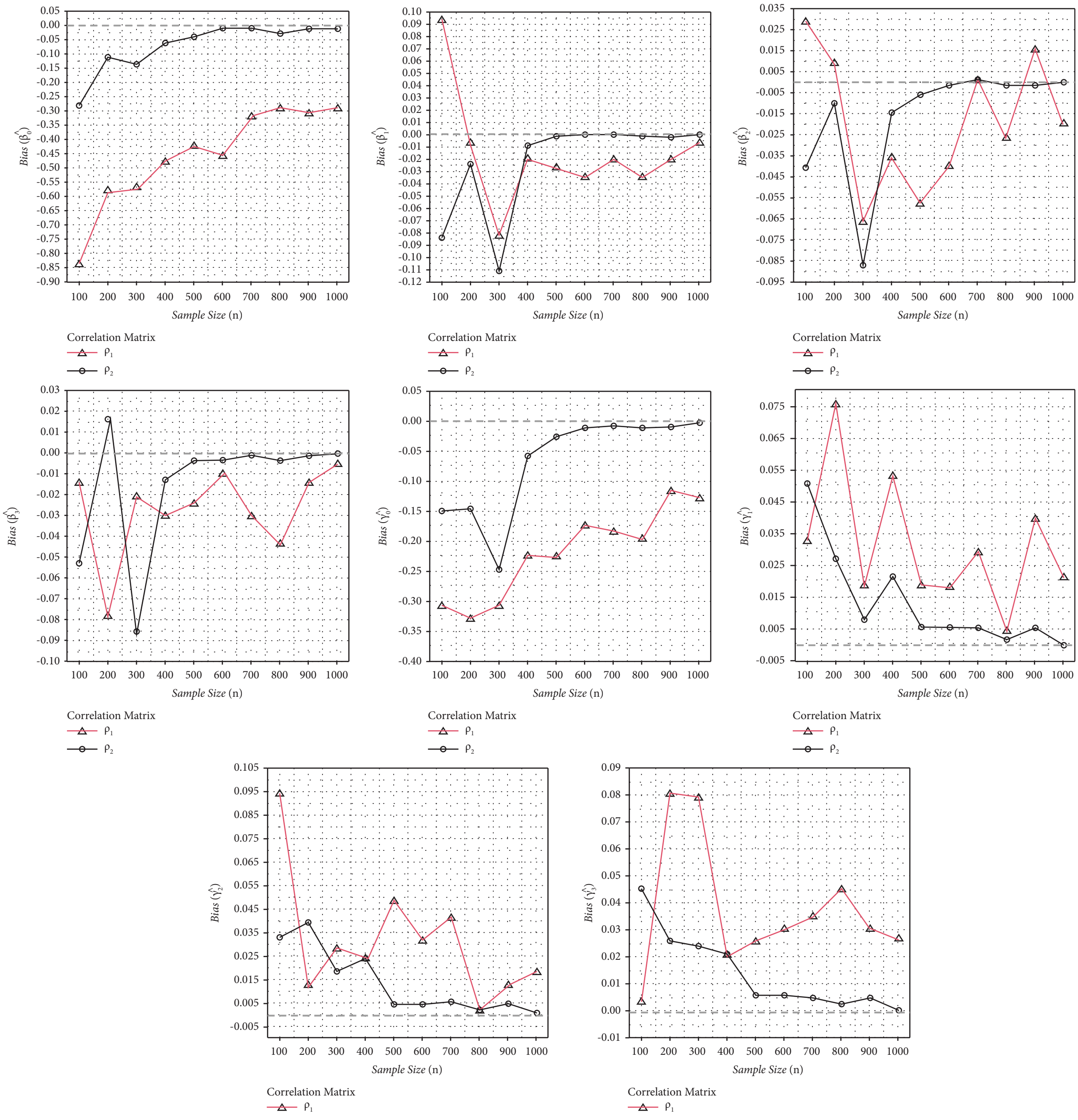

Figure 4: Continued.

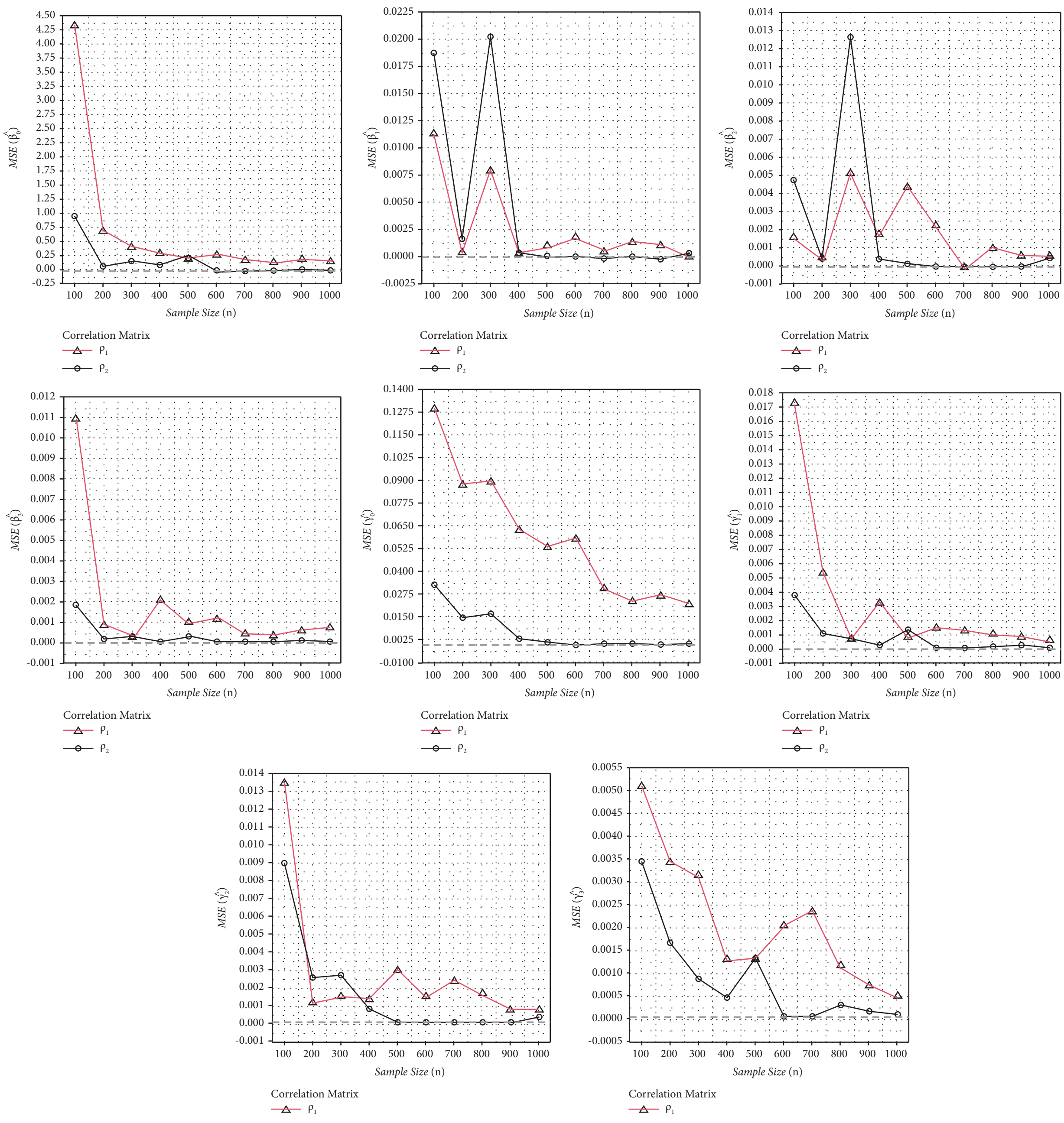

FIGURE 4: The biases and MSEs of the ZIPB count regression coefficients for the true parameters $\beta_2 = (0.1, 0.1, 0.1, 0.1)$ and $\gamma_2 = (0.2, 0.2, 0.2, 0.2)$ and the considered correlation matrices.

The overdispersion coefficient (θ) is statistically significant in the NB and ZINB regression models. However, it is not statistically significant in the zero-hurdle NB regression model ($p > 0.05$).

On the other hand, all coefficients except for γ_3 from the zero inflation coefficients are statistically significant in the ZIPB regression model.

Table 6 shows that the best-fitted regression model is the ZIPB regression model, according to all considered selection criteria. In this regard, we conclude that the ZIPB regression model is a good competitor to the ZIP, ZINB, zero hurdle-Poisson, and zero-hurdle-NB regression models.

6.2. Biochemistry Graduate Students Data. The second data set consists of 915 observations for each variable obtained by the pscl package in the *R* software. The data set is denoted bioChemists. Biochemistry graduate student (BGS) data refer to article production by graduate students in biochemistry Ph.D. programs. The BGS data set is anonymized,

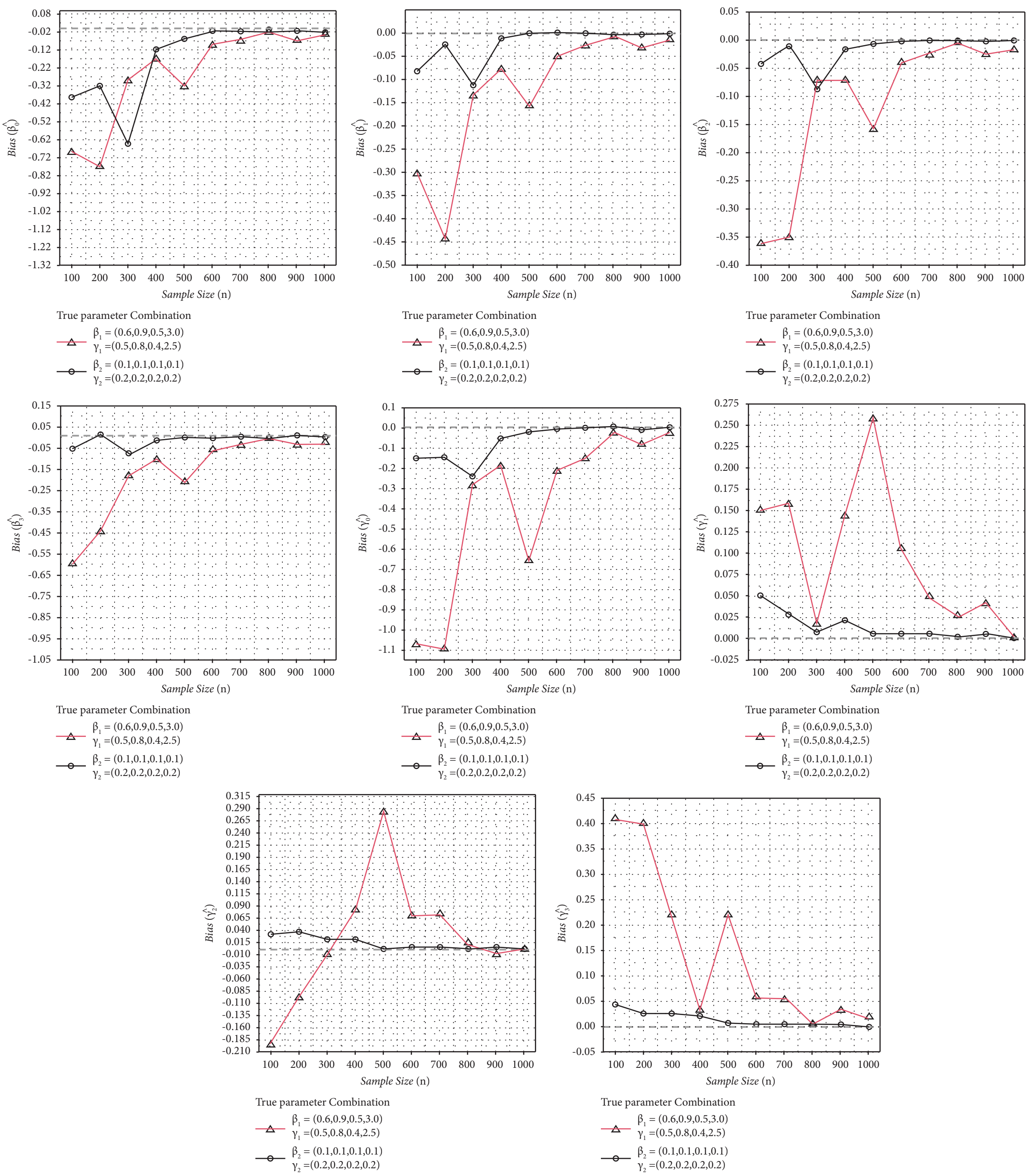

Figure 5: Continued.

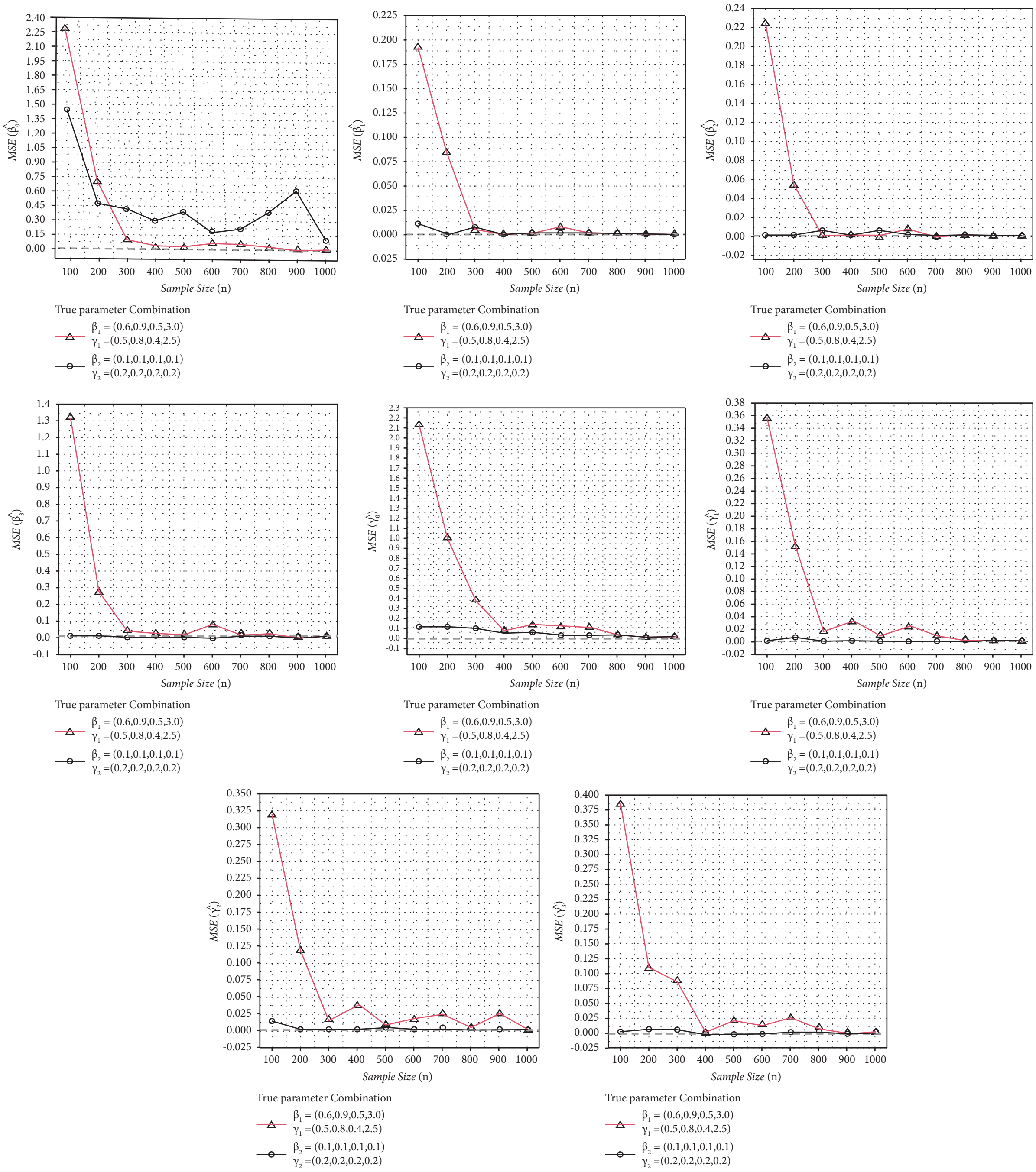

FIGURE 5: The biases and MSEs of the ZIPB count regression coefficients for the correlation matrix ρ_1 and for different values of the true parameters.

and it is open-access, available in [16]. For more details, see [17]. The considered covariates of the count regression model are given in Table 7.

Figure 8 illustrates the histogram of the response variable (the number of articles produced during last 3 years of Ph.D) for the BGS data set.

We analyze this data set with the Poisson, NB, PB, ZIP, ZINB, zero-hurdle-Poisson, zero-hurdle-NB, and ZIPB regression models. The MLEs and SEs of the fitted regression models are presented in Tables 8, 9, 10, and 11. We compare the fitted regression models via the statistics given in Table 12.

According to the results of the data analysis, the coefficient β_0 is statistically significant in the ZIP, zero-hurdle Poisson, and ZINB regression models ($p < 0.05$). The coefficient β_1 is statistically significant in the Poisson, PB, and NB regression models. In other words, the marriage status has an effect on the number of articles according to the

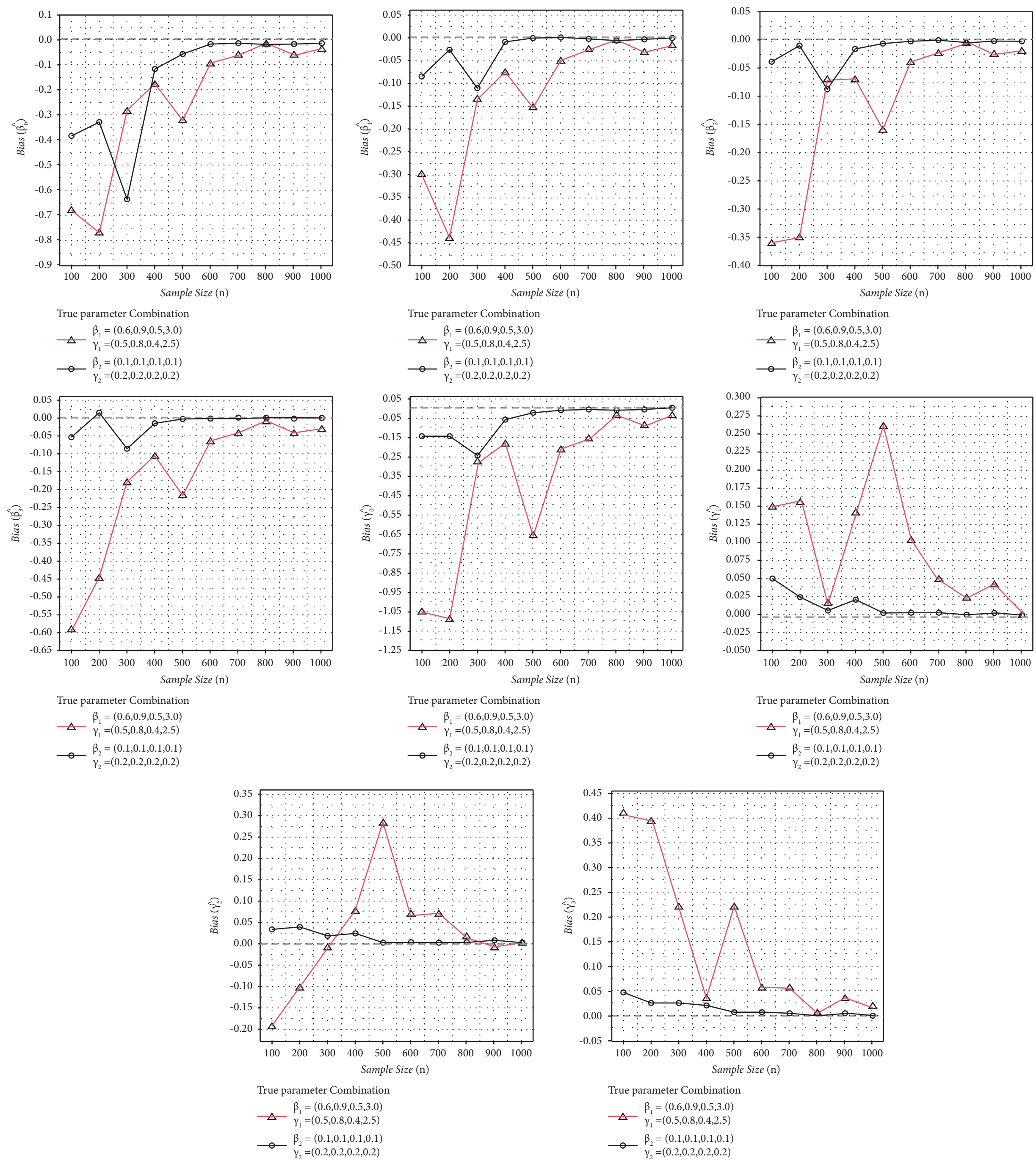

Figure 6: Continued.

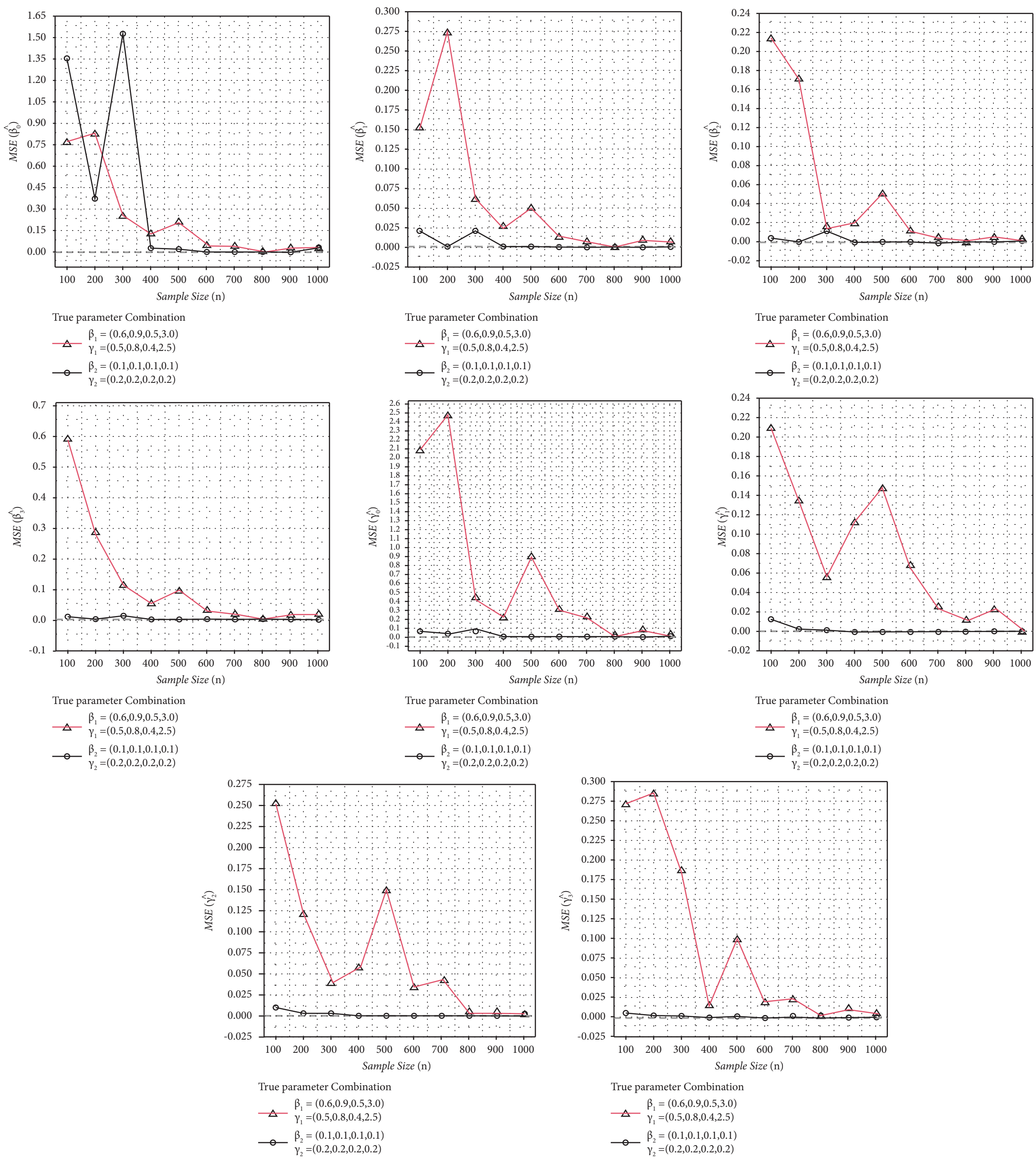

Figure 6: The biases and MSEs of the ZIPB count regression coefficients for the correlation matrix ρ_2 and for different values of the true parameters.

Table 1: The description of the AHS data.

y_i (doctorco)	"Number of consultations with a doctor or specian in the past 2 weeks"
x_{i1} (sex)	"Sex"
x_{i2} (age)	"Age"
x_{i3} (income)	"Income"
x_{i4} (illness)	"The number of illnesses in the past two weeks"
x_{i5} (actdays)	"The number of days of reduced activity in the past two weeks due to illness or injury"
x_{i6} (hscore)	"General health questionnaire score using Goldberg's method"

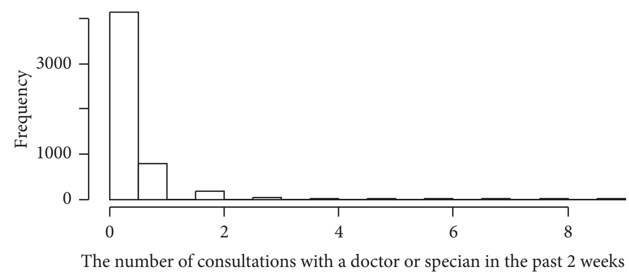

Figure 7: The histogram of the consultations with a doctor or specian in the past 2 weeks in the AHS data.

Table 2: The MLEs and the SEs of the Poisson and PB regression models for the AHS data set.

Covariates	Poisson					PB				
	MLE	S.E	95% CI	z value	p value	MLE	S.E	95% CI	z value	p value
β_0	−2.136	0.096	(−2.324, −1.948)	−22.271	<0.001	−2.274	0.108	(−2.485, −2.063)	−21.139	<0.001
β_1	0.188	0.055	(0.079, 0.296)	3.392	<0.001	0.232	0.063	(0.109, 0.355)	3.707	<0.001
β_2	0.514	0.135	(0.249, 0.778)	3.811	<0.001	0.592	0.151	(0.296, 0.889)	3.913	<0.001
β_3	−0.126	0.080	(−0.282, 0.031)	−1.565	0.11	−0.103	0.089	(−0.277, −0.071)	−1.157	0.250
β_4	0.198	0.018	(0.163, 0.232)	11.248	<0.001	0.218	0.021	(0.177, 0.258)	10.559	<0.001
β_5	0.128	0.005	(0.118, 0.137)	26.153	<0.001	0.138	0.006	(0.125, 0.150)	21.876	<0.001
β_6	0.031	0.010	(0.011, 0.05)	3.142	0.001	0.036	0.012	(0.013, 0.060)	3.004	0.002

Table 3: The MLEs and the SEs of the ZIP and ZIPB regression models for the AHS data set.

Covariates	ZIP					ZIPB				
	MLE	SE	95% CI	z value	p value	MLE	SE	95% CI	z value	p value
β_0	−0.605	0.139	(−0.878, −0.332)	−4.345	<0.001	−0.989	0.165	(−1.311, −0.666)	−6.007	<0.001
β_1	−0.027	0.071	(−0.165, 0.112)	−0.376	0.707	0.004	0.081	(−0.154, −0.162)	0.049	0.968
β_2	−0.127	0.175	(−0.470, 0.216)	−0.724	0.469	0.035	0.205	(−0.367, −0.436)	0.168	0.872
β_3	−0.150	0.105	(−0.356, 0.055)	−1.429	0.153	−0.121	0.117	(−0.350, −0.108)	−1.034	0.303
β_4	0.040	0.024	(−0.006, 0.087)	1.679	0.093	0.049	0.027	(−0.004, −7.102)	1.787	0.075
β_5	0.085	0.006	(−0.073, 0.096)	14.313	<0.001	0.106	0.007	(−0.091, −8.119)	14.810	<0.001
β_6	0.016	0.011	(−0.005, 0.037)	1.452	0.146	0.020	0.013	(−0.005, −5.045)	1.552	0.1212
Zero-inflation model coefficients										
γ_0	2.710	0.286	(−0.335, 1.907)	9.468	<0.001	2.585	0.362	(−1.874, −1.295)	7.131	<0.001
γ_1	−0.547	0.169	(−0.823, −0.153)	−3.243	0.001	−0.689	0.220	(−1.119, −1.258)	−3.135	0.001
γ_2	−2.651	0.421	(4.085, 16.907)	−6.306	<0.001	−2.903	0.564	(−4.008, −1.797)	−5.147	<0.001
γ_3	−0.028	0.234	(−20.569. −6.105)	−0.119	0.905	0.002	0.295	(−0.575, −0.579)	0.007	0.999
γ_4	−0.492	0.083	(−0.956. 0.082)	−5.901	<0.001	−0.749	0.138	(−1.020, −4.478)	−5.413	<0.001
γ_5	−1.283	0.255	(−0.819. −0.047)	−5.040	<0.001	−1.854	0.668	(−3.163, −7.544)	−2.774	0.005
γ_6	−0.109	0.039	(−0.688. 1.304)	−2.752	0.005	−0.135	0.058	(−0.248, −3.020)	−2.321	0.020

Poisson, PB, and NB regression models. "kid5" (β_2) has a statistically significant effect on "art" according to the Poisson, NB, PB, ZINB, and zero-hurdle Poisson regression models. "phd" (β_3) has not a statistically significant effect on "art" according to all the examined regression models. "ment" (β_4) is statistically significant in all the regression

Table 4: The MLEs and the SEs of the zero-hurdle-NB and ZINB regression models for the AHS data set.

Covariates	Zero-hurdle-NB					ZINB				
	MLE	SE	95% CI	z value	p value	MLE	SE	95% CI	z value	p value
β_0	−3.904	4.307	(−12.344, 4.537)	−0.906	0.365	−2.265	0.120	(−2.500, −2.029)	−18.830	<0.001
β_1	−0.018	0.157	(−0.325, 0.290)	−0.112	0.911	0.204	0.069	(0.06, 0.339)	2.943	0.003
β_2	−0.400	0.380	(−1.145, 0.345)	−1.052	0.293	0.561	0.167	(0.234, 0.887)	3.369	0.001
β_3	−0.365	0.221	(−0.798, 0.068)	−1.652	0.099	−0.152	0.097	(−0.343, −0.038)	−1.565	0.118
β_4	0.089	0.054	(−0.015, 0.194)	1.667	0.096	0.247	0.024	(0.201, 0.293)	10.494	<0.001
β_5	0.164	0.017	(0.130, 0.197)	9.509	<0.001	0.143	0.008	(0.127, 0.157)	18.635	<0.001
β_6	0.012	0.028	(−0.043, 0.066)	0.414	0.679	0.036	0.014	(−2.500, −2.029)	2.662	0.008
$\log\theta$	−3.409	4.445	(−12.122, 5.304)	−0.767	0.443	−0.051	0.096	(−0.240, 0.136)	−0.539	0.589
Zero-inflation model coefficients										
γ_0	−2.759	0.138	(−3.029, −2.487)	−19.955	<0.001	−2.115	3.471	(8.917, 4.688)	−0.609	0.542
γ_1	0.317	0.081	(−0.158, −0.475)	3.920	<0.001	−15.828	22.047	(−59.039, 27.383)	−0.718	0.473
γ_2	1.098	0.194	(−0.718, −1.477)	5.667	<0.001	−19.463	9.097	(−37.292, −1.633)	−2.140	0.032
γ_3	−0.010	0.112	(−0.229, −0.209)	−0.086	0.931	−15.076	8.308	(−31.358, 1.206)	−1.815	0.070
γ_4	0.286	0.027	(−0.232, −0.339)	10.429	<0.001	3.839	2.125	(−0.325, 8.004)	1.807	0.071
γ_5	0.159	0.012	(−0.136, −0.182)	13.639	<0.001	−1.296	1.213	(−3.672, 1.081)	−1.068	0.285
γ_6	0.061	0.017	(−0.027, −0.094)	3.542	<0.001	−0.528	0.388	(−1.287, 0.231)	−1.362	0.173

Table 5: The MLEs and the SEs of the zero-hurdle-Poisson and NB regression models for the AHS data set.

Covariates	Zero-hurdle-Poisson					NB				
	MLE	SE	95% CI	z value	p value	MLE	SE	95% CI	z value	p value
β_0	−0.515	0.166	(-0.841, −0.189)	−3.097	0.002	−2.35	0.116	(−2.577, −2.122)	−20.185	<0.001
β_1	−0.032	0.088	(−0.205, 0.140)	−0.369	0.712	0.250	0.069	(0.115, 0.383)	3.643	<0.001
β_2	−0.354	0.218	(−0.781, 0.072)	−1.628	0.104	0.649	0.165	(0.325, 0.972)	3.935	<0.001
β_3	−0.358	0.141	(−0.633, −0.0814)	−2.538	0.011	−0.097	0.097	(−0.286, 0.092)	−1.001	0.317
β_4	0.075	0.029	(0.018, 0.132)	2.595	0.009	0.230	0.022	(0.185, 0.273)	10.225	<0.001
β_5	0.115	0.007	(0.100, 0.128)	16.376	<0.001	0.145	0.007	(0.130, 0.159)	20.265	<0.001
β_6	0.006	0.014	(−0.021, 0.032)	0.405	0.685	0.038	0.013	(0.011, 0.064)	2.830	0.004
θ	—	—	—	—	—	0.927	0.086	(0.757, 1.097)	10.685	<0.001
Zero-inflation model coefficients										
γ_0	−2.759	0.138	(−3.029, −2.487)	−19.955	<0.001	—	—	—	—	—
γ_1	0.317	0.081	(−0.158, −0.475)	3.920	<0.001	—	—	—	—	—
γ_2	1.098	0.194	(−0.718, −1.477)	5.667	<0.001	—	—	—	—	—
γ_3	−0.010	0.112	(−0.229, −0.209)	−0.086	0.931	—	—	—	—	—
γ_4	0.286	0.027	(−0.232, −0.339)	10.429	<0.001	—	—	—	—	—
γ_5	0.159	0.012	(−0.136, −0.182)	13.639	<0.001	—	—	—	—	—
γ_6	0.061	0.017	(−0.027, −0.094)	3.542	<0.001	—	—	—	—	—

Table 6: Comparison criteria of the fitted regression models for the AHS dataset.

	-2ℓ	AIC	AICc	BIC
Poisson	6728.9	6742.9	6742.922	6754.906
NB	6412.118	6428.118	6428.146	6441.839
PB	6455.6	6469.6	6469.622	6481.606
ZIP	6398	6426	6426.081	6450.012
ZINB	6392	6422	6422.093	6447.728
Zero-hurdle-NB	6326	6356	6356.093	6381.728
Zero-hurdle-Poisson	6472	6500	6500.081	6524.012
ZIPB	6252.688	6280.688	6280.769	6304.7

models ($p < 0.05$). The overdispersion coefficient (θ) is statistically significant in the NB, ZINB, and zero-hurdle-NB regression models ($p < 0.05$). On the other hand, the coefficients γ_1 and γ_4 from zero inflation coefficients are statistically significant in the ZIPB regression model.

We observe that the best-fitted regression model is the ZINB regression model according to all examined selection criteria. Also, the ZIPB regression model is better than some well-known regression models, such as the Poisson, PB, ZIP, NB, zero-hurdle-Poisson, and zero-hurdle-NB models

Table 7: The description of the BGS data.

y_i (art)	"Count of articles produced during last 3 years of Ph.D."
x_{i1} (mar)	"Factor indicating marital status of student, with levels single and married"
x_{i2} (kid5)	"The number of children aged 5 or younger"
x_{i3} (phd)	"Prestige of Ph.D. department"
x_{i4} (ment)	"Count of articles produced by Ph.D. mentor during last 3 years"

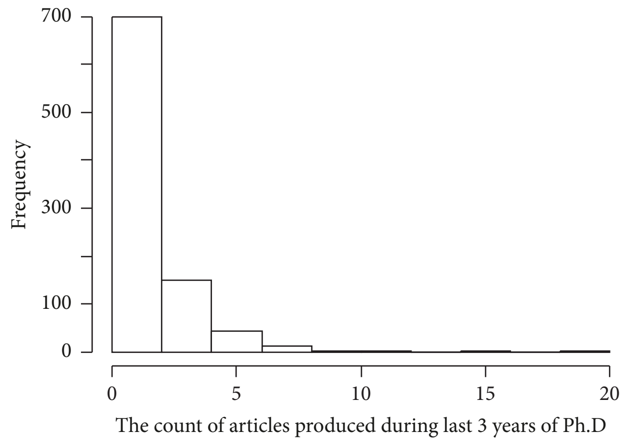

Figure 8: The number of the articles produced during last 3 years of Ph.D. in the BGS data set.

Table 8: The MLEs and SEs of the Poisson and PB regression models for the BGS data set.

Covariates	Poisson					PS				
	MLE	SE	95% CI	z value	p value	MLE	SE	95% CI	z value	p value
β_0	0.154	0.096	(−0.034, 0.342)	1.598	0.109	−0.071	0.187	(−0.437, 0.294)	−0.383	0.37
β_1	0.188	0.061	(0.068, 0.306)	3.091	0.001	0.180	0.084	(−0.014, 0.345)	2.135	0.04
β_2	−0.154	0.039	(−0.231, −0.076)	−3.912	<0.001	−0.144	0.054	(−0.249, 0.038)	−2.679	0.011
β_3	0.015	0.026	(−0.036, 0.067)	0.583	0.559	0.017	0.037	(−0.055, 0.090)	0.464	0.358
β_4	0.026	0.002	(0.022, 0.03)	13.237	<0.001	0.030	0.004	(−0.023, 0.037)	8.364	<0.001

Table 9: The MLEs and SEs of the ZIP and ZIPB regression models for the BGS data set.

Covariates	ZIP					ZIPB				
	MLE	SE	95% CI	z value	p value	MLE	SE	95% CI	z value	p value
β_0	0.512	0.116	(0.285, 0.738)	4.423	<0.001	0.152	0.197	(−0.234, 0.539)	0.772	0.296
β_1	0.133	0.071	(−0.005, −0.271)	1.882	0.059	0.117	0.088	(−0.054, 0.289)	1.336	0.163
β_2	−0.116	0.047	(−0.207, −0.024)	−2.485	0.013	−0.125	0.056	(−0.234, −0.014)	−2.212	0.034
β_3	−0.005	0.031	(−0.065, −0.055)	−0.167	0.867	0.001	0.039	(−0.075, 0.076)	0.017	0.398
β_4	0.019	0.002	(0.014, 0.023)	8.189	<0.001	0.027	0.004	(−0.019, 0.034)	7.163	<0.001
Zero-inflation model coefficients										
γ_0	−0.452	0.472	(−1.376, 0.472)	−0.959	0.338	2.881	1.827	(−0.699, 6.461)	1.577	0.115
γ_1	−0.382	0.308	(−0.984, 0.221)	−1.241	0.215	−2.237	1.030	(−4.256, −0.217)	−2.171	0.037
γ_2	0.190	0.192	(−0.186, 0.567)	0.989	0.323	0.676	0.545	(−0.392, 1.744)	1.240	0.184
γ_3	−0.005	0.143	(−0.284, 0.275)	−0.034	0.973	−0.172	0.362	(−0.881, 0.537)	−0.475	0.356
γ_4	−0.135	0.043	(−0.219, −0.050)	−3.126	0.002	−0.989	0.336	(−0.647, −0.329)	−2.942	0.005

Table 10: The MLEs and the SEs of the NB and zero-hurdle-Poisson regression models for the BGS data set.

Covariates	NB					Zero-hurdle-Poisson				
	MLE	SE	95% CI	z value	p value	MLE	SE	95% CI	z value	p value
β_0	0.114	0.129	(−0.138, 0.366)	0.884	0.377	0.530	0.116	(0.302, 0.756)	4.576	<0.001
β_1	0.180	0.082	(−0.020, 0.340)	2.205	0.027	0.128	0.072	(−0.013, −0.268)	1.769	0.077
β_2	−0.144	0.052	(−0.246, −0.042)	−2.774	0.006	−0.113	0.048	(−0.205, −0.019)	−2.362	0.018
β_3	0.016	0.036	(−0.054, 0.086)	0.449	0.653	−0.011	0.031	(−0.072, −0.050)	−0.350	0.726
β_4	0.030	0.003	(−0.023, 0.036)	9.347	<0.001	0.019	0.002	(0.015, 0.023)	8.553	<0.001
θ	2.198	0.259	(1.690, 2.705)	8.486	<0.001	—	—	—	—	—
Zero-inflation model coefficients										
γ_0	—	—	—	—	—	0.057	0.272	(−0.477, 0.590)	0.209	0.835
γ_1	—	—	—	—	—	0.366	0.179	(−0.015, 0.717)	2.047	0.041
γ_2	—	—	—	—	—	−0.247	0.108	(−0.458, 0.034)	−2.281	0.023
γ_3	—	—	—	—	—	0.025	0.079	(−0.130, 0.180)	0.316	0.752
γ_4	—	—	—	—	—	0.081	0.013	(−0.055, 0.106)	6.223	<0.001

Table 11: The MLEs and the SEs of the zero-hurdle-NB and ZINB regression models for the BGS data set.

Covariates	Zero-hurdle-NB					ZINB				
	MLE	SE	95% CI	z value	p value	MLE	SE	95% CI	z value	p value
β_0	0.192	0.191	(−0.183, 0.567)	1.004	0.316	0.294	0.135	(0.028, 0.558)	2.174	0.030
β_1	0.134	0.110	(−0.081, 0.349)	1.218	0.223	0.117	0.083	(0.045, 0.280)	1.408	0.159
β_2	−0.118	0.072	(−0.258, 0.022)	−1.642	0.101	−0.122	0.054	(2.226, −0.016)	−2.271	0.023
β_3	−0.004	0.049	(−0.099, 0.091)	−0.076	0.939	−0.001	0.036	(−0.072, 0.070)	−0.029	0.977
β_4	0.025	0.004	(−0.016, 0.033)	5.776	<0.001	0.026	0.004	(0.018, 0.032)	7.335	<0.001
$\log\theta$	0.545	0.225	(−0.105, 0.985)	2.429	0.015	0.942	0.134	(0.679, 1.203)	7.045	<0.001
Zero-inflation model coefficients										
γ_0	0.057	0.272	(−0.477, 0.590)	0.209	0.835	0.485	1.042	(−1.556, 2.526)	0.465	0.642
γ_1	0.366	0.179	(−0.015, 0.717)	2.047	0.041	−1.893	0.921	(−3.697, −0.088)	−2.056	0.040
γ_2	−0.247	0.108	(−0.458, 0.034)	−2.281	0.023	0.610	0.473	(−0.316, 1.536)	1.290	0.197
γ_3	0.025	0.079	(−0.130, 0.180)	0.316	0.752	−0.106	0.318	(−0.729, 0.518)	−0.332	0.740
γ_4	0.081	0.013	(−0.055, 0.106)	6.223	<0.001	−0.858	0.309	(−1.463, −0.253)	−2.781	0.005

Table 12: Selection criteria statistics of the fitted regression models for the BGS data set.

	-2ℓ	AIC	AICc	BIC
Poisson	3319.200	3329.200	3329.266	3334.007
NB	3130.700	3142.700	3142.793	3148.469
PB	3131.504	3141.504	3141.570	3146.311
ZIP	3223.601	3247.601	3247.947	3259.138
ZINB	3109.138	3135.138	3135.542	3147.636
Zero-hurdle-NB	3113.979	3139.979	3140.383	3152.477
Zero-hurdle-Poisson	3225.628	3249.628	3249.974	3261.165
ZIPB	3113.712	3137.712	3138.058	3149.249

according to all criteria. We conclude that the ZIPB regression model is an alternative to PB, ZINB, ZIP, zero-hurdle-Poisson, and zero-hurdle-NB regression models in the analysis of overdispersed data.

7. Conclusions

In the relevant literature, authors deal with the development of flexible discrete distributions in modeling overdispersed data. In this paper, we suggest a new zero-inflated regression model using the PB distribution. The ZIPB regression model was proposed as an alternative to some known count regression models, such as the Poisson, PB, NB, ZIP, ZINB, zero-hurdle-Poisson, and zero-hurdle-NB regression models. Two practical data examples were presented to assess the usefulness of this model and compare its performance with other competitors. The best-fitted regression model was the ZIPB model, according to all the criteria considered in the first data analysis. However, the ZINB regression model was the best-fitted regression model according to all considered criteria. The ZIPB regression model was the second well-fitted regression model. Although the ZINB regression model seems to be the best as a result of the second data set, we clearly observe that the

ZIPB regression model is superior to the zero-hurdle-NB regression model proposed as an alternative to the ZINB regression model. Sometimes, the zero-hurdle-NB regression model is known to be superior to the ZINB regression. In conclusion, we consider the proposed ZIPB regression model to be a good alternative to the ZINB and zero-hurdle-NB regression models and other zero-inflated regression models.

Data Availability

All the data used for this research are included in the manuscript.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors extend their appreciation to the Deputyship for Research and Innovation, Ministry of Education in Saudi Arabia, for funding this research work through project number 445-9-209.

References

[1] E. Altun, "A new zero-inflated regression model with application," *İstatistikçiler Dergisi: İstatistik ve Aktüerya*, vol. 11, no. 2, pp. 73–80, 2018.

[2] C. Tanış, H. Koç, and A. Pekgör, "A new zero–inflated discrete Lindley regression model," *Communications in Statistics-Theory and Methods*, vol. 53, no. 12, pp. 4252–4271, 2023.

[3] M. Denuit, X. Maréchal, S. Pitrebois, and J. F. Walhin, *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems*, John Wiley and Sons, Hoboken, New Jersey, USA, 2007.

[4] R. Shanker and H. Fesshaye, "On Poisson-Sujatha distribution and its applications to model count data from biological sciences," *Biometrics and Biostatistics International Journal*, vol. 3, no. 4, pp. 1–7, 2016.

[5] R. Shanker, "The discrete Poisson-aradhana distribution," *Turkiye Klinikleri Journal of Biostatistics*, vol. 9, no. 1, pp. 12–22, 2017.

[6] R. Shanker, K. K. Shukla, and T. A. Leonida, "A two-parameter Poisson-Akash distribution with properties and ap-plications," *International Journal of Probability and Statistics*, vol. 7, no. 4, pp. 114–123, 2018.

[7] E. Altun, "A new two-parameter discrete Poisson-generalized Lindley distribution with properties and applications to healthcare data sets," *Computational Statistics*, vol. 36, no. 4, pp. 2841–2861, 2021.

[8] J. M. Hilbe, *Negative Binomial Regression*, Cambridge University Press, Cambridge, UK, 2011.

[9] A. J. Lemonte, G. Moreno-Arenas, and F. Castellares, "Zero inflated bell regression models for count data," *Journal of Applied Statistics*, vol. 47, no. 2, pp. 265–286, 2020.

[10] R. R. M. Tajuddin, N. Ismail, K. Ibrahim, and S. A. A. Bakar, "A four-parameter negative binomial-Lindley distribution for modeling over and underdispersed count data with excess zeros," *Communications in Statistics-Theory and Methods*, vol. 51, no. 2, pp. 414–426, 2022.

[11] R. R. M. Tajuddin, N. Ismail, K. Ibrahim, and S. A. A. Bakar, "A new zero–one-inflated Poisson–Lindley distribution for modelling overdispersed count data," *Bulletin of the Malaysian Mathematical Sciences Society*, vol. 45, no. S1, pp. 21–35, 2022.

[12] E. Altun, "A new one-parameter discrete distribution with associated regression and integer-valued autoregressive models," *Mathematica slovaca*, vol. 70, no. 4, pp. 979–994, 2020.

[13] A. M. Abd-Elrahman, "Utilizing ordered statistics in lifetime distributions production: a new lifetime distribution and applications," *Journal of probability and statistical science*, vol. 11, no. 2, pp. 153–164, 2013.

[14] N. Johnson, S. Kotz, and A. Kemp, *Univariate Discrete Distributions*, John Wiley and Sons, New York, NY, USA, 2nd edition, 1992.

[15] 2023, https://search.r-project.org/CRAN/refmans/faraway/html/dvisits.html.

[16] J. S. Long, "The origins of sex differences in science," *Social Forces*, vol. 68, no. 4, pp. 1297–1316, 1990.

[17] J. S. Long, *Regression Models for Categorical and Limited Dependent Variables*, Sage, Thousand Oaks, California, 1997.