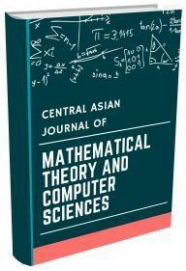

CENTRAL ASIAN JOURNAL OF MATHEMATICAL THEORY AND COMPUTER SCIENCES

https://cajmtcs.centralasianstudies.org/index.php/CAJMTCS

Volume: 06 Issue: 03 | July 2025 ISSN: 2660-5309

Article

Beta Principal Component Regression Model with Application

Rammah Oday Hassan

1. Assistant Lecturer of Statistic at Administration and Economics of Babylon University, Iraq
* Correspondence: rimah.dunbl@uobabylon.edu.iq

Abstract: The main goal of regression analysis is to estimate the effect relationship between independent variables and the dependent variable. Consequently, the type of response variable data determines the type of regression model to be used. If the dependent variable data is continuous and represents proportions confined between (0, 1), the Beta regression model is considered a good choice for representing such a relationship. However, at times, the Beta regression model encounters issues that make the estimated relationship unstable. This instability is due to the presence of some econometric problems that cause the estimators to be inaccurate. One such issue is multicollinearity. This problem arises when there is a strong correlation between the explanatory variables, which particularly affects the estimators of the parameters (β) by reducing their accuracy. Multicollinearity leads to an unusual inflation of parameter variances, making the estimators less reliable. In this paper, we will present a new method that integrates the Beta regression model with Principal Component Regression to develop a hybrid regression model. This hybrid model will be more suitable for estimation in the presence of multicollinearity issues. Simulation examples and real data are used to evaluate the performance of the proposed method in comparison with existing methods.

Keywords: Beta Regression, Principal Component Regression, Multicollinearity, Beta Principal Component Regression

Citation: Hassan, R. O. Beta Principal Component Regression Model with Application. Central Asian Journal of Mathematical Theory and Computer Sciences 2025, 6(3), 591-600.

Received: 03th Mar 2025
Revised: 11th Apr 2025
Accepted: 24th May 2025
Published: 17th Jun 2025

1. Introduction

The beta regression model (BReg) is an active tool for evaluating the relationship between continuous response variables that are bound within a limited range, typically between 0 and 1, and a set of independent variables. Since the seminal work of Ferrari, S. L. P., & Cribari-Neto, F. B Reg has become very popular in various application sciences, for example : Medical sciences, nature science. medical science, econometric science. The Beta regression model is considered one of the generalized linear regression models [1], [2]. This model is particularly suitable for cases where the response variable follows a Beta distribution, and the data for this variable are continuous and defined within the open interval (0,1). When classical regression is applied to such data, which follow a Beta distribution, the model loses its ability to predict and generalize effectively. Moreover, when relative data are transformed into discrete values, the predicted and expected values may fall outside the closed interval. Beta regression is adaptable due to the Beta distribution's ability to assume a multitude of forms, which enables it to accommodate a diverse array of data patterns. Nevertheless, it experiences difficulties when the number of predictors exceeds the number of observations, or when the predictors (independent variables) are highly correlated (multicollinearity) [3]. To overcome these real challenges, the researchers proposed an extension of principal component regression (PCReg) regression specifically with Beta regression model. This extension allows us for a good treatment of multicollinearity problem.in this paper,we will proposed a new statistical

technique that combines principal component regression (PCreg) with beta linear regression to address the problem of multicollinearity between predictor variables.This method transforms the original correlated variables into a set of uncorrelated components, which can then be used as predictors in regression models, enhancing prediction accuracy and interpretability [4]. The following sections of this paper are structured as follows.in section 2 the multicollinearity problem that affect the accuracy of independent variales model. In section 3 Simple introduction of beta regression. In section 4 proposed beta principal component regression model. We illustrate the effectiveness of our proposed method through both simulation analyses and empirical data. with brief conclusion shown in section 6 [5].

The Multicollinearity Problem

Multicollinearity problem in multiple beta regression appear from high correlations between explanatory variables, complicating the evaluation of individual variable effects. This problem negatively affect model stability, accuracy and difficult of the interpretation for real relationships between explanatory variables and response variable. The multicollinearity problem with the beta regression model to makes the beta regression model unreliable [6]. Where, Coefficients maybe not reflect the real relationship between explanatory variables and the response variable. The size or direction of an effect. The multicollinearity may be wrong, also the multicollinearity problem raises the variance of coefficient estimates and makes it difficult to determine whether a explanatory variables are statistically significant. Multicollinearity in beta regression models has a significant impact on the variance of maximum likelihood (ML) estimation [7], [8], [9], [10]. Because of, presence of a multicollinearity problem can provide us over fitting, in which the beta regression model captures a lot of noise in the data instead of the main relationship. Multicollinearity presents substantial issues in beta regression modeling, compromising the stability and interpretability of results.

2. Materials and Methods

Beta regression model

beta distribution

The beta distribution is a probability distribution used to model random variables with values ranging from 0 to 1. This makes it particularly handy when modeling proportions or rates. The beta distribution is characterized by two parameters: a and b. The notation $a > 0$ and $b > 0$ indicate that these parameters must be greater than zero. The values of a and b influence the distribution's mean, variance, and overall look. Below, the probability density function (PDF) of the Beta distribution can be presented as follows:

$$f(y,a,b) = \frac{\Gamma(a+b)}{\Gamma a\,\Gamma b} \quad y^{a-1}(1-y)^{b-1} \qquad 0 < y < 1 \qquad (1)$$

$\Gamma(x) is\ named\ Gamma\ function$, $\mathrm{B(a,b)} is\ named\ beta\ function$, according this information the equation (1) rewrite as following:

$$f(y,a,b) = \frac{1}{\mathrm{B(a,b)}} \quad y^{a-1}(1-y)^{b-1} \quad 0 < y < 1 \qquad (2)$$

where $\mathrm{B(a,b)} = \frac{\Gamma a\,\Gamma b}{\Gamma(a+b)}$, therefore $\frac{1}{\mathrm{B(a,b)}} = \frac{\Gamma(a+b)}{\Gamma a\,\Gamma b}$, $y\ belong\ to\ open\ interval\ (0,1). a > 0$ and $b > 0$,

$\Gamma(y) = (y-1)!$. According to Equation 1, the Beta distribution is one of the continuous distributions that has a mean ($\frac{a}{a+b}$) and variance ($\frac{ab}{(a+b)^2(a+b+1)}$)(Damgaard, C. F., & Irvine, K. M. (2019)). We shall reparametrize because in regression models, parameterization can assist make the model more interpretable and computationally efficient. It enables for parameter transformations, which can improve convergence properties during estimate. The reparametrization can be by adding precision parameter($\emptyset > 0$). Therefore , $\emptyset = a + b, then\ the\ \mu = \frac{a}{a+b} \Rightarrow_{,} \mu = \frac{a}{\emptyset} \Rightarrow_{,} a = \emptyset\mu$, we can find b via precision parameter $\emptyset$ as following:

$a = (a+b)\mu = a - a\mu = b\mu \Rightarrow_{,} b = \frac{a(1-\mu)}{\mu}\, b = \frac{\emptyset\mu(1-\mu)}{\mu} \Rightarrow_{,} b = (1-\mu)\emptyset$,according the value $\emptyset$,the mean is equal $\mu = \frac{a}{\emptyset}$, $\sigma^2 = \frac{\mu(1-\mu)}{1+\emptyset}$. Therefore, the probability density function (PDF) of beta distribution is became as following:

$$f(y,a,b)=\frac{\Gamma(\emptyset)}{\Gamma\emptyset\mu\,\Gamma(1-\mu)\emptyset}\quad y^{\emptyset\mu-1}(1-y)^{(1-\mu)\emptyset-1}\qquad 0<y<1\qquad (3)$$

$(\emptyset,\mu>0,)$(Cribari-Neto, F., & Zeileis, A. (2010).).

Beta regression model

Beta regression is a statistical approach used to describe response variables that are continuous and limited between 0 and 1. It is especially suitable for proportions and rates. The model is based on the assumption that the response variable follows a beta distribution, which allows for greater flexibility in modeling data of diverse shapes and variances. The beta regression model often includes establishing a link function to relate the mean of the response variable to the predictor variables.

Let the observed response variable is $y_i \in (0,1), i=1,2,3,\ldots,n$ where the response variable (y_i) is follow a beta distribution: $y_i \sim Beta(\mu_i\emptyset,(1-\mu_i)\emptyset$. The link function $g_{(.)}$is to related the response variable (y_i) and information matrix X

$$g_{(\mu_i)}=x_i^T\beta \qquad (4)$$

where $g_{(.)}$ is link function between expected value of the dependent variable $\mu_i=E(y_i)$ and linear independent variables $x_i^T\beta\,.x_i^T$ is the vector of independent variables(Firinguetti, L., et al(2024)), where $x_i^T=(1,x_{i1},x_{i2},\ldots\ldots x_{ik})$ and β is vector of known regression parameters. We will used logit function $(\log\frac{\mu_i}{1-\mu_i})$ as link function by some special derivative $\mu_i=\frac{\exp(x_i^T\beta)}{1-\exp(x_i^T\beta)}$. Finally, link function (logit function) is a key component of beta regression models. It aids in converting proportions into a format that can be studied using traditional statistical methods, making it an invaluable tool for authors dealing with rates and proportions. By using an appropriate link function, a Beta regression model can be constructed to estimate the parameters (β) as well as the precision parameter$(\emptyset)$.

Beta principal component regression model

Often, we encounter a set of challenges when estimating the parameters of a regression model. One of these challenges is the problem of multicollinearity. To obtain efficient estimators, it is necessary to use methods that are more robust to this problem. One of the proposed methods is the Principal Component Analysis (PCA) method. To get the beta principal component regression model is a statistical technique that combines principal component analysis with beta regression modeling, allowing for the reduction of multicollinearity among independent variables while improving prediction accuracy and interpretability. This method is particularly useful in situations where the number of predictors exceeds the number of observations, as it helps to extract the most important features from a dataset and mitigate potential overfitting issues. By transforming the original predictors into a smaller set of uncorrelated variables, As in the following formula:

$$D=X\Lambda \qquad (5)$$

where D is the principal components matrix of order $(n\times q)$ and Λ is an orthogonal matrix of standardized eigenvectors corresponding to the eigenvalues of the information matrix (X^tX) , with order $(q*q)$ and elements (θ_{ij}) $i=1,\ldots\ldots n$, and its columns $(\Lambda_j$)j=1,2,…..q , This matrix Λ diagonalizes the information matrix(Greenacre, M., et al (2022)), assuming that the eigenvalues of the information matrix (X^tX), $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_q$. To construct a regression model, as shown in Equation (5), based on the principal component D, we express the dependent variable y_i as a function of the orthogonal principal components rather than the intercorrelated independent variables X, Since the matrix Λ is orthogonal, we have the property $\Lambda\Lambda^t=I$,where I is the identity matrix . Thus, the equation (5) can be rearranged to find the original variables in terms of the principal component matrix as follows:

$$X=D\Lambda^T \qquad (6)$$

Therefore, the link function $g_{(\mu_i)}=X\beta$ is became $g_{(\mu_i)}=D\Lambda^T\beta$ .Let $\Lambda^T\beta=\theta$,from this information the link function is $g_{(\mu_i^*)}=D\theta$, thus the $Y=D\theta+U$. To estimation of the parameters θ and ∅ via The maximum likelihood method is $l(\theta,\emptyset)=\sum_{i=1}^n l_i(\mu_i^*,\emptyset)$ (7)

$l_i(\mu_i^*,\emptyset)=ln\,\Gamma(\emptyset)-\ln(\Gamma\emptyset\mu_i^*)-\Gamma(1-\mu_i^*)\emptyset+(\mu_i^*\emptyset-1)\ln y_i+[(1-\mu_i^*)\emptyset]\ln(1-y_i)$ (8)

To estimate the parameters of the model, the equation above can be solved using Fisher scoring technique is one method of a numerical solution. Through a series of derivations using numerical methods, the parameters of a model can be estimated using the following equation.

$$\hat{\theta}_{ML} = [(D\Lambda^T)^T W D\Lambda^T]^{-1}(D\Lambda^T)^T Wz \quad (9)$$

where $\hat{\theta}_{ML}$ is estimated to parameters of principal component, D is principal components matrix, W is weighted matrix. To obtain the original estimates for the regression coefficients of the beta principal component regression model $\hat{\beta}$. it can be achieved through feedback as follows:

$$\Lambda^T\hat{\beta} = \hat{\theta} \Rightarrow \hat{\beta} = \Lambda\hat{\theta} \quad (10)$$

The beta principal component regression enhances model performance and provides clearer insights into the relationships between variables. This technique not only streamlines the modeling process but also enables researchers to identify underlying patterns and relationships that may have been obscured by noise in the data.Based on the information provided above, we will design a reliable and efficient algorithm to estimate the parameters of the model under investigation. Our algorithm is executed for 10,000 iterations, with the initial 1,000 discarded as burn-in.

3. Result

Simulation approach

In the subsequent section, the efficacy of the proposed method is evaluated via comprehensive simulation studies. We conduct a comparative analysis of our beta principal component regression approach (BPCReg) against the Beta regression model as delineated in Ferrari, S. L. P., & Cribari-Neto, F. referred to as 'BReg. The two methods are evaluated based on the mean absolute error(MAE) that compute as follows $MAE = \frac{1}{r}\sum_{i=1}^{n}|y_i - \hat{y}_i|$. The median of mean absolute deviations (MMAD) are also used, where MMAD that compute as follows: $MMAD = median\ (mean|x_i^T\hat{\beta} - x_i^T\beta^{true}|$. The data in this simulation approach are generated as following

1. The data of dependent variable $y_i$is simulated from beta distribution with vector of mean $\mu = (\mu_i), i = 1,2, \dots n$, where $\mu_i = \frac{\exp(x_i^T\beta)}{1-\exp(x_i^T\beta)}$.
2. The values of precision parameter $\emptyset$, is identification by the set {1,6}, $\emptyset\ \epsilon${1,6}.
3. Therefore $y_i \sim beta(a = \mu_i\emptyset_i, b = 1-\mu_i)\emptyset_i))$
4. The independent variables are simulated from beta distribution standard normal distribution with mean zero and variance one, where $x_{i\sim N(0,1)}$.
5. To achieve varying levels of multicollinearity among independent variables. We will use three levels of correlation for the independent variables, with the first being the low correlation level ($\rho = 0.18\ and\ \rho = 0.30$), The Intermediate level correlation level($\rho = 0.40\ and\ \rho = 0.55$) and The high level correlation level($\rho = 0.86\ and\ \rho = 0.96$) [11], [12], [13].
6. The two simulation examples are used with initial parameters for first example $\beta = (0.2374, 0.1062, 0.5217,\ 0.5173,\ 0.3146, 0.3422, 0.4202)^T$ and second example $\beta = \begin{pmatrix} 0.2907, 0.5213, 0.3671, 0.4491, 0.3974, 0.000, 0.0931, 0.2041, 0.000, \\ 0.0884, 0.3020, 0.000 \end{pmatrix}^T$. both the two simulation examples contain of intercept term $\beta = 0$ and $\sum_{j=1}^{p}\beta_j$. $p = 7,12$ respectively

Two simulated examples are analyzed:

Simulation example one

n this simulated example the number of independent variables are seven variables (p=7). From the structure of our simulation above, we obtund the following results

Table 1. The results of MAE and MMAD for our proposed method (BPCReg) and BReg with precision parameter $\emptyset = 1$.

	levels of multicollinearity

Comparison Methods	$\rho = 0.18$	$\rho = 0.30$	$\rho = 0.40$	$\rho = 0.55$	$\rho = 0.86$	$\rho = 0.96$
	MMAD (MAE)	MMAD (MAE)	MMAD (MAE)	MMAD (MAE)	MMAD (MAE)	MMAD (MAE)
N=50						
BReg	1.8621 (1.5412)	1.9542 (1.7436)	1.4520 (1.3457)	0.8910 (0.8146)	0.8752 (0.8457)	**0.8691** (0.7574)
BPCReg	1.5563 (1.4267)	1.6132 (1.6424)	1.2101 (1.1245)	0.8579 (0.7945)	0.8634 (0.7972)	**0.8372** (0.7158)
N=100						
BReg	1.6321 (1.5485)	1.6015 (1.5647)	1.1374 (1.1348)	0.8603 (0.7124)	**0.8634** (0.7248)	**0.8346** (0.7024)
BPCReg	1.3527 (1.3457)	1.3162 (1.4571)	1.1195 (1.0124)	0.8424 (0. 7117)	**0.8396** (0.7124)	**0.7822** (0.6974)
N=150						
BReg	1.4526 (1.3240)	1.5722 (1.3471)	1.0101 (1.0065)	**0.9520** (0.6781)	0.8590 (0.6926)	0.8111 (0.7541)
BPCReg	1.2025 (1.1457)	1.2514 (1.2158)	0.9532 (0.9817)	0.9431 (0.6672)	**0.8327** (0.6542)	**0.7816** (0.7213)
N=200						
BReg	1.3135 (1.0548)	1.4612 (1.1624)	1.1842 (0.9723)	0.8745 (0.6461)	**0.8586** (0.6357)	0.7815 (0.7121)
BPCReg	1.1823 (0.9472)	1.2423 (0.9871)	1.1595 (0.9647)	0.8521 (0.6217)	0.8332 (0.6157)	0.7712 (0.6871)
N=250						
BReg	1.2866 (0.9256)	1.3623 (0.9637)	1.1291 (0.9568)	0.8567 (0.6127)	0.8372 (0.6034)	0.7381 (0.5417)
BPCReg	1.1180 (0.9124)	1.1299 (0.9724)	1.0943 (0.9428)	0.8245 (0.6049)	0.8116 (0.5817)	0.7496 (0.5216)

Note: In the parentheses are M**AE**

Table 1. presents a synopsis of the MAE and MMAD for the two methods in the comparison. It is shows that from Table 1 that the performance of BPCReg appears very good compared to the BReg method. In general, the MMAD and MAE computed by our proposed method is much smaller than MMAD and MAE computed by BReg method [14]. As we observe from the results in the table above, when the correlation coefficients increase, the values of the Two criteria measures (MMAD), (MAE) decrease across all sample sizes. We observe that our proposed method has outperformed the another method across all studied scenarios precision parameter $\emptyset = 1$.

Table 2. The results of MAE and MMAD for our proposed method (BPCReg) and BReg with precision parameter $\emptyset = 6$.

levels of multicollinearity						
Comparison Methods	$\rho = 0.18$	$\rho = 0.30$	$\rho = 0.40$	$\rho = 0.55$	$\rho = 0.86$	$\rho = 0.96$
	MMAD (MAE)	MMAD (MAE)	MMAD (MAE)	MMAD (MAE)	MMAD (MAE)	MMAD (MAE)
N=50						
BReg	1.2395 (1.2182)	1.1078 (1.0567)	1.0927 (1.0021)	**0.6835** (0.5674)	0.537 8 (0.5248)	0.5272 (0.5127)
BPCReg	1.2051 (1.1548)	1.0032 (0.9587)	1.0122 (0.9642)	**0.6484** (0.5428)	0.4874 (0.5064)	0.5143 (0.5028)
N=100						
BReg	1.3132 (0.1154)	**0.9874** (0.9245)	0.9232 (0.9347)	0.8224 (0.5127)	**0.6415** (0.4541)	0.5752 (0.4572)
BPCReg	1.1143	0. 8632	0.9126	0.8013	**0.6262**	0.565 6

	(0.1064)	(0.9142)	(0.9108)	(0.5024)	(0.4317)	(0.4361)
N=150						
BReg	**0.9861** (0.9457)	0.8942 (0.9057)	0.8934 (0.8928)	0.7596 (0.4829)	**.05381** (0.4288)	0.5334 (0.4127)
BPCReg	0.9637 (0.9542)	**0.7552** (0.8275)	0.8235 (0.8561)	0.7017 (0.4687)	**.05113** (0.4187)	0.5222 (0.4067)
N=200						
BReg	0.8145 (0.7854)	0.7525 (0.8140)	0.7354 (0.8246)	**0.5817** (0.4381)	**0.5302** (0.4029)	0.5523 (0.4005)
BPCReg	0.7533 (0.6781)	0.7135 (0.7951)	0.6212 (0.7861)	**0.5364** (0.4218)	**0.5145** (0.3829)	0.5227 (0.3981)
	N=250					
BReg	0.7236 (0.6582)	0.652 5 (0.6942)	0.6352 (0.7124)	**0.5198** (0.4125)	**0.4734** (0.3719)	0.4821 (0.3762)
BPCReg	0.6741 (0.6425)	0.5423 (0.6572)	0.6521 (0.6827)	0.4842 (0.3648)	0.4213 (0.3527)	0.4136 (0.3595)

Note: In the parentheses are M**AE**

Table 2. presents a synopsis of the MAE and MMAD for the two methods in the comparison. It is shows that from Table 2 that the performance of BPCReg appears very good compared to the BReg method . In general, the MMAD and MAE computed by our proposed method is much smaller than MMAD and MAE computed by BReg method [15], [16]. As we observe from the results in the table above, when the correlation coefficients increase, the values of the Two criteria measures (MMAD), (MAE) decrease across all sample sizes. We observe that our proposed method has outperformed the another method across all studied scenarios with precision parameter $\emptyset = 6$.

Simulation example two

In this simulated example the number of independent variables are nine variables (p=12). From the structure of our simulation above, we obtund the following results

Table 3. The results of MAE and MMAD for our proposed method (BPCReg) and BReg with precision parameter $\emptyset = 1$.

levels of multicollinearity						
Comparison Methods	$\rho = 0.50$	$\rho = 0.75$	$\rho = 0.85$	$\rho = 0.90$	$\rho = 0.95$	$\rho = 0.99$
	MMAD	MMAD	MMAD	MMAD	MMAD	MMAD
N=50						
BReg	0.8725 (0.7865)	0.8467 (0.7924)	0.8246 (0.7692)	0.8178 (0.8563)	0.8052 (0.7567)	0.7924 (0.7659)
BPCReg	0.8514 (0.7496)	0.8134 (0.7689)	0.8125 (0.7538)	0.8100 (0.8267)	0.7954 (0.7457)	0.7642 (0.7466)
N=100						
BReg	**0.5827** (0.5748)	**0.5881** (0.5542)	**0.5742** (0.5506)	0.5685 (0.5147)	**0.5620** (0.5296)	**0.5913** (0.5792)
BPCReg	**.05342** (0.5627)	**.05172** (0.5279)	**.05013** (0.5274)	0.4962 (0.5095)	0.4843 (0.5129)	**0.4460** (0.5597)
N=150						
BReg	0.8071 (0.5469)	0.7645 (0.5197)	0.7428 (0.5196)	0.6133 (04927)	0.5941 (0.5067)	0.5644 (0.4387)
BPCReg	0.7723 (0.5356)	0.7591 (0.4983)	0.6121 (0.4837)	0.5971 (0.4728)	0.5783 (0.4837)	0.5681 (0.4264)
N=200						
BReg	**0.5743** (0.5249)	**0.5543** (0.4834)	**0.5617** (0.4682)	0.5543 (0.4692)	0.5620 (0.4672)	0.5254 (0.4186)

BPCReg	**.05221** (0.5149)	**.04942** (0.4751)	**.04928** (0.4583)	0.4724 (0.4398)	0.4538 (0.4463)	0.4362 (0.3861)
N=250						
BReg	0.7945 (0.4864)	0.7755 (0.4711)	0.7546 (0.4276)	0.6744 (0.4167)	0.5748 (0.4291)	0.5529 (0.3672)
BPCReg	0.7436 (0.4638)	0.7282 (0.4567)	0.7243 (0.4189)	0.5822 (0.3951)	0.5732 (0.4062)	0.5346 (0.3514)

Note: In the parentheses are M**AE**

Table .3. presents a synopsis of the MAE and MMAD for the two methods in the comparison. It is shows that from Table 3 that the performance of BPCReg appears very good compared to the BReg method . In general, the MMAD and MAE computed by our proposed method is much smaller than MMAD and MAE computed by BReg method. As we observe from the results in the table above, when the correlation coefficients increase, the values of the Two criteria measures (MMAD), (MAE) decrease across all sample sizes [17]. We observe that our proposed method has outperformed the another method across all studied scenarios with precision parameter $\emptyset = 1$.

Table 4. The results of MAE and MMAD for our proposed method (BPCReg) and BReg with precision parameter $\emptyset = 6$.

levels of multicollinearity						
Comparison Methods	$\rho = 0.50$	$\rho = 0.75$	$\rho = 0.85$	$\rho = 0.90$	$\rho = 0.95$	$\rho = 0.99$
	MMAD	MMAD	MMAD	MMAD	MMAD	MMAD
BReg	**0.5483** (0.4567)	**0.5317** (0.4395)	**0.4237** (0.4058)	0.4208 (0.3767)	0.4098 (0.3978)	0.4045 (0.3471)
BPCReg	**0.5145** (0.4387)	**0.5226** (0.4189)	**0.4672** (0.3987)	0.4596 (0.3498)	0.4079 (0.3762)	0.3824 (0.3275)
N=100						
BReg	0.7814 (0.4167)	0.7561 (0.4068)	0.6789 (0.3794)	0.6522 (0.3128)	0.5639 (0.3542)	0.5377 (0.3149)
BPCReg	0.7295 (0.3853)	0.7202 (0.3892)	0.6703 (0.3591)	0.5384 (0.3059)	0.5581 (0.3351)	0.5147 (0.3019)
N=150						
BReg	**0.5348** (0.3647)	**0.5173** (0.3728)	**0.5056** (0.3381)	0.4016 (0.2972)	0.5297 (0.3187)	0.3924 (0.2897)
BPCReg	**0.5049** (0.3429)	**0.4832** (0.3517)	**0.4714** (0.3195)	0.4334 (0.2875)	0.3846 (0.3057)	0.3643 (0.2758)
N=200						
BReg	0.6437 (0.3259)	0.6243 (0.3498)	0.5425 (0.2978)	0.5246 (0.2761)	0.5169 (0.2954)	0.4329 (0.2637)
BPCReg	0.5227 (0.3128)	0.5864 (0.3384)	0.5015 (0.2792)	0.4924 (0.2697)	0.4577 (0.2873)	0.4253 (0.2548)
N=250						
BReg	**0.4838** (0.2972)	0.4264 (0.3275)	0.4655 (0.2781)	0.3827 (0.2458)	0.3733 (0.2691)	0.3571 (0.2497)
BPCReg	**0.4552** (0.2876)	0.4142 (0.3157)	0.4165 (0.2483)	0.3618 (0.2285)	0.3370 (0.2429)	0.3227 (0.2453)

Note: In the parentheses are M**AE**

Table 4. presents a synopsis of the MAE and MMAD for the two methods in the comparison. It is shows that from Table 4 that the performance of BPCReg appears very good compared to the BReg method . In general, the MMAD and MAE computed by our proposed method is much smaller than MMAD and MAE computed by BReg method [18], [19]. As we observe from the results in the table above, when the correlation coefficients increase, the values of the Two criteria measures (MMAD), (MAE) decrease

across all sample sizes. We observe that our proposed method has outperformed the another method across all studied scenarios with precision parameter $\emptyset = 6$.

Real Data

This study will concentrate on a medical phenomenon represented by herpes disease. Herpes, a widespread viral ailment, is triggered by the herpes simplex virus (HSV), which consists of two primary types: HSV-1, responsible for cold sores (oral herpes), and HSV-2, which usually results in sores in the genital region (genital herpes). Direct contact with the skin or bodily fluids of an infected individual facilitates the virus's transmission [20]. The our data is collected from Al-Diwaniyah Hospital with sample size 127 observation . This disease can be identified by an increase in the percentage of the (Immunoglobulin G) IgG index. Therefore, IgG index is represented the response variable (y is IgG index)is influenced by a set of independent variables, which are: x_1 :Packed Cell Volume (PCV), x_2 : Erythrocyte Sedimentation Rate (ESR) , x_3 : Platelets (PLT) , x_4 : Procalcitonin (PCT), x_5 : Mean Platelet Volume (MPV) Mean Platelet Volume and x_6 : Serum Creatinine (**S.Creatinine**). Before starting the data analysis using the proposed method, it is essential to check whether the data suffers from the problem of multicollinearity or not. In the current study, we will focus on the simplest measure, which is finding the correlation matrix between the independent variables.

Table 5. Shows the correlation matrix between the independent variables

Variables	PCV	ESR	PLT	PCT	MPV	S.Creatinine
PCV	1	0.231	0.056	0.152	0.006	0.121
ESR		1	0.547	0.894	0.258	0.793
PLT			1	0.247	0.125	0.381
PCT				1	0.527	0.142
MPV					1	0.491
S.Creatinine						1

From the correlation matrix results, we observe a high and clear correlation between some independent variables. The peak correlation coefficient between the two variables (ESR,PCT) reached (0.894). A high correlation between independent variables is considered a preliminary test for the problem of multicollinearity. By observing the results of the correlation matrix among our independent variables, we can determine the presence of multicollinearity in the study model. The estimated parameters for our proposed method and the other method can be presented in the table 5 below.

Table 6. Prameters estimates for our proposed method (BPCReg) and other method

The variables	BReg	BPCReg
Intercept	1.6421	0.5341
PCV	1.0062	0.2457
ESR	0.8654	0.2543
PLT	0.5672	0.5314
PCT	0.7678	0.4861
MPV	-0.0267	-1.5647
S.Creatinine	-0.5654	-0.0083

From the estimated parameter values above for both methods, we observe that there are both positive and negative effects on the response variable (IgG). We will rely on these estimated parameters to calculate the mean squared error (MSE) for the real data (Table 6).

Table 7. The values of (MSE) for the our proposed method BPCReg and BReg

Comparison Methods	MSE
BReg	1.9315
BPCReg	0.6824

From the results shown above, we observe that the mean squared error (MSE) calculated using our proposed method is significantly lower than the MSE obtained using the other comparison method. Based on this result, we can conclude that our proposed method also demonstrates high efficiency with real data, even in the presence of the multicollinearity problem.(Table 7)

4. Discussion

The results from both simulation and real data analysis strongly demonstrate the effectiveness of the Beta Principal Component Regression (BPCReg) model in addressing multicollinearity and enhancing prediction accuracy. Across various levels of correlation and sample sizes, BPCReg consistently yielded lower Mean Absolute Error (MAE) and Median of Mean Absolute Deviations (MMAD) compared to the standard Beta Regression (BReg) model. These improvements indicate that transforming the original correlated predictors into uncorrelated principal components helps to stabilize parameter estimates and improve model performance. In real data analysis using herpes-related medical data, BPCReg also outperformed BReg, evidenced by a substantially lower Mean Squared Error (MSE). These findings imply that BPCReg offers both robustness and generalizability, making it a suitable method in domains prone to multicollinearity, such as medical or econometric research. Nevertheless, one limitation of this approach lies in the interpretability of principal components, which may lack direct real-world meaning. Future research could explore integrating regularization techniques or Bayesian frameworks with BPCReg to further enhance its flexibility and estimation accuracy in more complex or high-dimensional datasets.

5. Conclusion

The structure and type of the response variable data determine the appropriate model type that will provide us with good estimators and models with high predictive power. When the data of the response variable consists of percentages, conventional regression models will fail to produce efficient estimators. To overcome this issue, a Beta regression model can be used. In most cases, regression models suffer from some standard problems, and it is difficult to find good estimators in the presence of these issues. One of the most serious standard problems is multicollinearity. To overcome this issue, multicollinearity can be addressed by using certain treatment methods. One of the methods used is Principal Component Regression (PCR). In this research, a combination of the Beta regression model with Principal Component Regression was applied to develop a robust model that effectively addresses the multicollinearity problem. We observe that our proposed model demonstrates very high efficiency in estimating the Beta regression model, even in the presence of perfect or semi-perfect multicollinearity. This model exhibits a high capability for generalization and flexibility in estimating Beta regression models from it captures the complex relationships through the independent variables. The results confirm that the our proposed method (BPCReg) is effective and robust when applied with different data.

Regression models should be carefully selected to match the available data in order to avoid obtaining misleading results that do not accurately represent the phenomenon being studied. When we encounter a econometrics problem in the model being studied, the appropriate method should be employed to overcome this issue. Therefore, we recommend using our proposed approach with Beta regression models in the presence of multicollinearity problems. We recommended extended our proposed method with regularization method to improve ability and generalizability.

REFERENCES

[1] F. H. H. Alhusseini and M. H. Odah, "Principal component regression for tobit model and purchases of gold," in Proc. 10th Int. Manag. Conf., Bucharest, Romania, vol. 10, pp. 491–500, 2016.

[2] W. F. Massy, "Principal components regression in exploratory statistical research," J. Amer. Stat. Assoc., vol. 60, no. 309, pp. 234–256, 1965.
[3] M. Greenacre, P. J. Groenen, T. Hastie, A. I. d'Enza, A. Markos, and E. Tuzhilina, "Principal component analysis," Nat. Rev. Methods Primers, vol. 2, no. 1, p. 100, 2022.
[4] W. F. Massy, "Principal components regression in exploratory statistical research," J. Amer. Stat. Assoc., vol. 60, no. 309, pp. 234–256, 1965.
[5] A. Junaid, A. Khan, A. Alrumayh, F. M. Alghamdi, E. Hussam, H. M. Aljohani, and A. Alrashidi, "Modified two parameter ridge estimator for beta regression model," J. Radiat. Res. Appl. Sci., vol. 17, no. 2, p. 100905, 2024.
[6] S. H. Mahmood, "Estimating models and evaluating their efficiency under multicollinearity in multiple linear regression: A comparative study," Zanco J. Human Sci., vol. 28, no. 5, pp. 264–277, 2024.
[7] A. Kalnins, "Multicollinearity: How common factors cause Type 1 errors in multivariate regression," Strateg. Manag. J., vol. 39, no. 8, pp. 2362–2385, 2018.
[8] M. R. Abonazel, H. A. Said, E. Tag-Eldin, S. Abdel-Rahman, and I. G. Khattab, "Using beta regression modeling in medical sciences: a comparative study," Commun. Math. Biol. Neurosci., Article ID, 2023.
[9] F. Bertrand, N. Meyer, K. El Bayed, I. J. Namer, and M. Maumy-Bertrand, "Regression beta PLS," J. Soc. Fr. Stat., vol. 154, no. 3, pp. 143–159, 2013.
[10] M. R. Abonazel, H. A. Said, E. Tag-Eldin, S. Abdel-Rahman, and I. G. Khattab, "Using beta regression modeling in medical sciences: a comparative study," Commun. Math. Biol. Neurosci., Article ID, 2023.
[11] E. A. Geissinger, C. L. Khoo, I. C. Richmond, S. J. Faulkner, and D. C. Schneider, "A case for beta regression in the natural sciences," Ecosphere, vol. 13, no. 2, p. e3940, 2022.
[12] S. L. P. Ferrari and F. Cribari-Neto, "Beta regression for modelling rates and proportions," J. Appl. Stat., vol. 31, no. 7, pp. 799–815, 2004.
[13] M. R. Abonazel, H. A. Said, E. Tag-Eldin, S. Abdel-Rahman, and I. G. Khattab, "Using beta regression modeling in medical sciences: a comparative study," Commun. Math. Biol. Neurosci., Article ID, 2023.
[14] R. A. Belaghi, Y. Asar, and R. Larsson, "Improved shrinkage estimators in the beta regression model with application in econometric and educational data," Stat. Papers, vol. 64, no. 6, pp. 1891–1912, 2023.
[15] M. D. Ismaiel and A. D. Ahmed, "Comparison between estimates of beta regression model using MSE," J. Econ. Admin. Sci., vol. 30, no. 143, pp. 523–535, 2024.
[16] Q. A. Owoyemi and A. Bolakale, "Comparative analysis of some linear predictive models in the presence of multicollinearity," Int. J. Adv. Stat. Probab., vol. 11, no. 1, 2023.
[17] H. Ali, C. N. Akanihu, and J. Felix, "Investigating the parameters of the beta distribution," World J. Adv. Res. Rev., vol. 19, no. 1, pp. 815–830, 2023.
[18] C. F. Damgaard and K. M. Irvine, "Using the beta distribution to analyse plant cover data," J. Ecol., vol. 107, no. 6, pp. 2747–2759, 2019.
[19] F. Cribari-Neto and A. Zeileis, "Beta regression in R," J. Stat. Softw., vol. 34, pp. 1–24, 2010.
[20] L. Firinguetti, M. González-Navarrete, and R. Machaca-Aguilar, "Shrinkage estimators for beta regression models," arXiv preprint, arXiv:2406.18047, 2024.