Mathematics and Statistics 10(4): 833-843, 2022
DOI: 10.13189/ms.2022.100414 http://www.hrpub.org

Generalized Biased Estimator for Beta Regression Model: Simulation and Application

Rasha A. Farghali, Samah M. Abo-El-Hadid*

Department of Mathematics, Insurance and Applied Statistics, Faculty of Commerce and Business Administration, Helwan University, Cairo, Egypt

Received May 9, 2022; Revised June 7, 2022; Accepted July 11, 2022

Cite This Paper in the Following Citation Styles

(a): *[1] Rasha A. Farghali, Samah M. Abo-El-Hadid , "Generalized Biased Estimator for Beta Regression Model: Simulation and Application," Mathematics and Statistics, Vol. 10, No. 4, pp. 833 - 843, 2022. DOI: 10.13189/ms.2022.100414.*

(b): *Rasha A. Farghali, Samah M. Abo-El-Hadid (2022). Generalized Biased Estimator for Beta Regression Model: Simulation and Application. Mathematics and Statistics, 10(4), 833 - 843. DOI: 10.13189/ms.2022.100414.*

Abstract Beta regression model is used for modeling proportions measured on a continuous scale; its parameters are estimated with the maximum likelihood method. Classical regression models, such as linear regression model and nonlinear regression models like logistic regression are not suitable for such situations. As in linear regression model, the independent variables are assumed to be uncorrelated if this assumption is not met, then the multicollinearity appears. Multicollinearity problem means that there is a near dependency between the independent variables. Biased estimators are commonly used for correcting the multicollinearity problem. In this study, we propose a generalized biased estimator for correcting multicollinearity in beta regression that is generalize beta ridge regression estimator (GBRRE). The performance of the proposed generalized biased estimator is evaluated theoretically via the matrix mean squared errors and the scalar mean squared errors; and practically using a Monte Carlo simulation study. The simulation results show that the optimal shrinkage estimator is K1 and the worst one is K2. Also, the proposed generalized estimator is applied to a real data set of pre-university education students in Egypt during the academic year (2018/2019) and we found the application results agree with the simulation results. Finally based on the results of the simulation study and the application the performance of the suggested generalized biased estimator is better than maximum likelihood estimators.

Keywords Beta Regression, Ridge Estimator, Multicollinearity, Generalized Ridge Regression

1. Introduction

Statistical modeling of continuous proportions has received close attention in the last few years. Some examples of proportions measured on a continuous scale include the fraction of income contributed to a retirement fund, the proportion of weekly hours spent on work-related activities, the proportion of income spent on food, the poverty rate, etc. see Ünlüa and Aktaş [39]. Classical regression models, such as linear regression model and nonlinear regression models like logistic regression are not suitable for such situations. For modeling continuous proportions several strategies have been proposed that are observed to be related to other variables. Beta regression models have been studied by Kieschnick and McCullough [32], Ferrari and Cribari-Neto [27], Espinheira, Ferrari and Cribari-Neto [22,23], Cribari-Neto and Zeileis [20]. Multicollinearity problem means that there is a near dependency between the independent variables. It was primarily studied in linear regression models when the ordinary least squares (OLS) method is used to estimate the regression parameters, $\hat{\beta}_{OLS}$ are unbiased estimators with minimum variances but their performance becomes inadequate when multicollinearity problem occurs. So that, the estimates $\hat{\beta}_{OLS}$ will be unbiased but have large variances, inflated confidence intervals and theoretically important variables become insignificant in testing hypotheses see Belsley [18]. Biased estimators are introduced to correct the multicollinearity problem in linear regression model as ridge regression estimator by Hoerl and Kennard [28,29], Eledum and Awadallah [21]. Liu and Liu-type estimators by Liu [33,34,35], Alheety and Kibria [9], Awwad et al. [17] among others. These estimators are also introduced to correct multicollinearity in nonlinear regression models as: binomial logistic regression see Asar [14], Asar and Genç [15,16], multinomial logistic regression see Abonazel and Farghali [1], gamma regression see Algamal [7], Amin et al. [12], Amin et al. [13], Akram et al. [6], Akram et al. [4], and Poisson regression see Amin et al., [10,11] among others.

Recently, the multicollinearity problem is considered in the beta regression model see Karlsson et al., [31], Qasim et al., [37], Abonazel and Taha [2], Algamal and Abonazel [8]. Motivated by the work of Hoerl and Kennard [28,29], Bhat and Vidya [19], Farghali [26], Farghali et al., [25] and Akarm et al., [5]. This paper aims to introduce a generalized biased estimator to deal with multicollinearity in beta regression model. The proposed estimator is a generalized estimator that includes: the maximum likelihood estimators (MLE) and the beta ridge regression estimator (BRRE). We theoretically and numerically assess the performance of the proposed estimator under some efficient parameters. The rest of the paper is arranged as follows: The beta regression model and the proposed estimator and mean square error (MSE) properties are well defined in Section 2. Estimation methods for selection of generalize optimal shrinkage parameter are explained in Section 3. Mont Carlo Simulation study and its results are presented in Section 4. The advantage of our recommended estimators is demonstrated by analyzing real data application in Section 5 and concluding remarks are given in Section 6.

2. Methodology

2.1. Beta Regression Model

The beta distribution with parameters ϑ and δ, denoted by $B_e(\vartheta,\delta)$, has the density function:

$$f(y;\,\theta,\delta)\ =\frac{\Gamma(\vartheta+\delta)}{\Gamma(\vartheta)\,\Gamma(\delta)}\ y^{\vartheta-1}\ (1\ -\ y)^{\delta-1}, \qquad (1)$$

where $0<y<1$, $\Gamma(\cdot)$ denotes the gamma function and $(0\ <\ \vartheta\ <\ 1\ and\ \delta\ >\ 0)$. $E(y)=\frac{\vartheta}{\vartheta+\delta}$ and $V(y)=\frac{\vartheta\delta}{(\vartheta+\delta)^2(\vartheta+\delta+1)}$, respectively. For suggesting a regression model of response variable that has beta distribution, Ferrari and Cribari-Neto [27] defined a re-parameterization based on eq.(1), they assumed that: $\mu=\frac{\vartheta}{\vartheta+\delta}$ and $\phi=\vartheta+\delta$, so that, $\vartheta=\mu\phi$ and $\delta=\phi(1-\mu)$, then by substituting in eq.(1) with the new values of (ϑ,δ) the density function in eq.(1) with the new parameters $(\mu\ and\ \phi)$, $0\ <\ \mu\ <\ 1\,,\phi\ >\ 0$, was as follows:

$$f(y;\,\mu,\phi)\ =\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\,\Gamma(\phi(1-\mu)}\ y^{\mu\phi-1}\ (1\ -\ y)^{\phi(1-\mu)-1}$$
$$0<y<1 \qquad (2)$$

Where, μ is the mean of the response variable and ϕ is the precision parameter. $E(y)=\mu$ and $V(y)=\frac{\mu(1-\mu)}{(\phi+1)}$, respectively. For a random sample $y_1,y_2,\ldots,y_n$, if $y_i{\sim}B_e(\mu_i,\phi)$ with density function:

$$f(y_i;\,\mu_i,\phi)\ =\frac{\Gamma(\phi)}{\Gamma(\mu_i\phi)\,\Gamma(\phi(1-\mu_i)}\ y_i^{\mu_i\phi-1}\ (1\ -\ y_i)^{\phi(1-\mu_i)-1}$$
$$0<y_i<1\,, i=1,2,\ldots,n \qquad (3)$$

let $g(.)$ be a strictly monotonic and twice differentiable link function that maps $(0,1)\ into\ \mathcal{R}$, then the beta regression model can be defined such that:

$$g(\mu_i)=\eta_i=\sum\nolimits_{j=1}^{p^*}\beta_j x_{ij}\ ,$$
$$(\text{where, j}=1,2,\ldots,p^*\ \text{and}\ p^*=p+1) \qquad (4)$$

Where, x_{ij} (pre-determined) is the measurement of the jth explanatory variables for the ith observation, $j=1,2,\ldots,p^*, i=1,2,\ldots,n$. $\beta_1,\beta_2,\ldots,\beta_{p^*}$ (unknown) regression parameters. Several link functions may be used for fitting the beta regression model such as: logit, probit, log–log, complementary log–log and Cauchy link functions. Ferrari and Cribari-Neto [27] suggested the logit link function, i.e.

$$log\left(\frac{\mu_i}{1-\mu_i}\right)=\sum_{j=1}^{p^*}\beta_j x_{ij} \qquad (5)$$

$$\mu_i=\frac{e^{\sum_{j=1}^{p^*}\beta_j x_{ij}}}{1+e^{\sum_{j=1}^{p^*}\beta_j x_{ij}}} \qquad (6)$$

Where, μ_i is the mean response function. Since η_i depends on β_j and the mean response μ_i is a function of η_i , then, the means $\mu_1,\mu_2,\ldots,\mu_n$ are functions of β_j , $,j=1,2,\ldots,p^*, i=1,2,\ldots,n$. From equations (3) and (6) the log likelihood function is as follows:

$$log\,f(y_i;\,\mu_i,\phi)=$$
$$\sum_{i=1}^{n}log[\Gamma\phi-log[\Gamma(\mu_i\phi)]-log[\Gamma\phi(1-\mu_i)]]+$$
$$(\mu_i\phi-1)\,log\,y_i+[\phi(1-\mu_i)-1]\,log(y_i-1) \qquad (7)$$

The maximum likelihood method is used to estimate the regression parameters $\beta_j, j=1,2,\ldots,p^*$ and the dispersion parameter ϕ . The maximum likelihood estimates (MLE) can be found by setting the first derivative of (7) to zero, but the obtained system of equations is nonlinear. The solution to this nonlinear system of equations is found by applying numerical methods as: Newton-Raphson method and Fisher score method. The score function of the parameters $\beta_j, j=1,2,\ldots,p^*$ is presented as:

$$S(\beta)=\phi\,X'M(Y^*-\mu^*) \qquad (8)$$

Where, M is a $(n\times n)$ diagonal matrix its elements are $\left(\frac{1}{g'(\mu_1)},\frac{1}{g'(\mu_2)},\ldots,\frac{1}{g'(\mu_n)}\right)$, Y^* is a $(n\times 1)$ vector,

$y_i^* = log\frac{y_i}{1-y_i}$, $i = 1,2,\dots,n$, $\mu^*$is a $(n \times 1)$ vector:

$$\mu_i^* = \gamma\left[\frac{e^{\sum_{j=1}^{p^*}\beta_j x_{ij}}}{1+e^{\sum_{j=1}^{p^*}\beta_j x_{ij}}}\phi\right] - \gamma\phi\left[1-\frac{e^{\sum_{j=1}^{p^*}\beta_j x_{ij}}}{1+e^{\sum_{j=1}^{p^*}\beta_j x_{ij}}}\right]$$

where $\gamma(.)$ indicates the digamma function. The Fisher scoring algorithm or the iterative reweighted least square (IRWLS) algorithm can be used for estimating the parameters $\beta_j, j = 1,2,\dots,p^*$, Consider $\hat{\beta}^r$ as the maximum likelihood estimator with r iterations, then β^r will be as follows:

$$\hat{\beta}^{r+1} = \hat{\beta}^r + \left(I(\beta^r)\right)^{-1} S(\beta^r) \quad (9)$$

where $S(\beta^r)$ as defined in Eq (7) and r = 0, 1, 2, … are the iterations which are implemented until convergence is reached and $I(\beta^r) = -E\left(\frac{\partial^2 l(\beta)}{\partial\beta_j\beta_h}\right)$ is the Fisher information matrix and it is evaluated at β^r, the iterations will end when the difference

$\hat{\beta}^r - \hat{\beta}^{r+1} < \epsilon$, ϵ a certain small value usually equals 10^{-6} . With some simplifications, the last form of estimation algorithm is then associated with iterative reweighted least squares as:

$$\hat{\beta}_{MLE} = \left(X'\widehat{W}X\right)^{-1}\left(X'\widehat{W}z\right) \quad (10)$$

Where, $\left(X'\widehat{W}X\right)$ is a $p^* \times p^*$ Fisher information matrix, $\widehat{W}$ is a $(n \times n)$ diagonal matrix with its main element equals:

$$\widehat{w}_i = \widehat{\phi}_\iota\left[\gamma'\left(\frac{e^{\sum_{j=1}^{p^*}\hat{\beta}_j x_{ij}}}{1+e^{\sum_{j=1}^{p^*}\hat{\beta}_j x_{ij}}}\widehat{\phi}_\iota\right) + \gamma'\left(1\frac{e^{\sum_{j=1}^{p^*}\hat{\beta}_j x_{ij}}}{1+e^{\sum_{j=1}^{p^*}\hat{\beta}_j x_{ij}}}\right)\widehat{\phi}_\iota\right]\frac{1}{\left(g'(\hat{\mu}_i)\right)^2} \quad (11)$$

and $z = \sum_{j=1}^{p^*}\hat{\beta}_j x_{ij} + \widehat{W}^{-1}\widehat{M}(Y^* - \mu^*), \gamma'(.)$ indicates the trigamma function. $\widehat{W}, \widehat{M}$ are the estimated matrices of W, M. Ferrari and Cribari-Neto [27] proved that Under the usual regularity conditions for maximum likelihood estimation with large sample size, $\left(\hat{\beta}\right) \sim N\left(\beta, \frac{1}{\phi}\left(X'\widehat{W}X\right)^{-1}\right)$ approximately. The scalar mean squared error $(SMSE)$ of $\hat{\beta}_{MLE}$ is obtained as:

$$SMSE\left(\hat{\beta}_{MLE}\right) = \frac{1}{\phi}tr(X'\widehat{W}X)^{-1} = \frac{1}{\phi}\sum_{j=1}^{p^*}\frac{1}{\lambda_j}, \quad (12)$$

where λ_j is the j^{th} eigenvalue of the estimated information matrix $(X'\widehat{W}X)$, $j = 1,2,\dots,p^*$. When the multicollinearity problem exists λ_j, $j = 1,2,\dots,p^*$ will tend to zero so that the values of $(X'\widehat{W}X)^{-1}$ are inflated.

2.2. Proposed Generalized Estimator

Regularization methods using penalization helps to deal with the multicollinearity problem by putting constraints on the values of the estimated parameters, thus, the entries of the variance-covariance matrix are also reduced. Hoerl and Kennard [28,29] introduced one of the oldest penalization methods for linear models which is ridge penalty, known as L_2 penalized regression, the ridge regression estimates $\hat{\beta}_{RR}$ are obtained by minimizing the ridge regression criterion. It fixes the multicollinearity problem by adding a single shrinkage parameter k to the main diagonal of the information matrix so that, $\left\|\hat{\beta}_{RR}\right\| < \left\|\hat{\beta}_{OLS}\right\|$. This led to biased estimates with smaller values and smaller standard errors than those of $\hat{\beta}_{OLS}$. Some penalized estimators are introduced to deal with multicollinearity in beta regression model which are: Karlsson et al., [31] introduced beta Liu estimator, Abonazel and Taha [2], Qasim et al., [37] introduced beta ridge regression estimator (BRRE), Algamal and Abonazel [8] introduced beta Liu-type estimator, Akram et al. [4] introduced beta modified ridge type estimator. Based on the work of Hoerl and Kennard [28,29], Abonazel and Taha [2] introduced a beta ridge regression estimator as follows:

$$\hat{\beta}_{BRR} = \left(X'\widehat{W}X + k\,I\right)^{-1}X'\widehat{W}z \ , \ k > 0 \quad (13)$$

With $\left\|\hat{\beta}_{BRR}\right\| < \left\|\hat{\beta}_{MLE}\right\|$. Also, Qasim et al., [47] introduced ridge regression as follows:

$$\hat{\beta}_{BRR} = \left(X'\widehat{W}X + k\,I\right)^{-1}(X'\widehat{W}X)\hat{\beta}_{MLE} \ , \ k > 0 \quad (14)$$

Where k is the shrinkage parameter and I is a $p^* \times p^*$ identity matrix. Using simulation experiment they proved that all proposed ridge regression estimators with different shrinkage parameters outperform MLE, that encouraged the simulation results for the linear regression model of Hoerl et al. [28,29], Issa [30] and Kibria [32] among others.

This study aims to generalize the beta ridge regression estimator (BRRE). The proposed estimator is defined as:

$$\hat{\beta}_{GBRRE} = \left(X'\widehat{W}X + K\right)^{-1}\left(X'\widehat{W}X\right)\hat{\beta}_{MLE}. \quad (15)$$

where $K = diag\{k_j\}; \ k_j > 0, \ j = 1,2,\dots,p^*$. The asymptotic covariance matrix is calculated as follows:

$$Cov\left(\hat{\beta}_{GBRRE}\right) = \frac{1}{\phi}\left(X'\widehat{W}X + K\right)^{-1}\left(X'\widehat{W}X\right)\left(X'\widehat{W}X + K\right)^{-1} \quad (16)$$

Lemma 1: The suggested generalized biased estimator in (15), represents a general case, as it is easy to see that: $\lim_{K\to O}\hat{\beta}_{GBRRE} \to \hat{\beta}_{MLE}$;

When $k_1 = k_2 = \cdots = k_{p^*} = k$, then

$\hat{\beta}_{GBRRE} = \left(X'\widehat{W}X + k\mathrm{I}\right)^{-1}\left(X'\widehat{W}X\right)\hat{\beta}_{MLE} = \hat{\beta}_{BRRE}$ that is the beta ridge regression estimator as in Eq(14)

2.3. The Matrix and the Scalar Mean Squared Error Properties of the Proposed Estimator

The matrix mean squared error (MMSE) of an estimator $\widetilde{\varpi}$ of the regression paramter ϖ is defined as:

$$\begin{aligned}\mathrm{MMSE}(\widetilde{\varpi}) &= E(\widetilde{\varpi} - \varpi)(\widetilde{\varpi} - \varpi)' \\ &= Cov(\widetilde{\varpi}) + bias(\widetilde{\varpi})bias(\widetilde{\varpi})' \quad (17)\end{aligned}$$

Where, $Cov(\widetilde{\varpi})$ represents the covariance matrix of $\widetilde{\varpi}$ and $bias(\widetilde{\varpi}) = E(\widetilde{\varpi}) - \varpi$. The scalar mean squared error (SMSE) is another criterion to evaluate the goodness of an estimator and it is defined as:

$$\text{SMSE}(\varpi) = tr[Cov(\varpi)] + bias(\varpi)'bias(\varpi) \quad (18)$$

To present the precise form of the (MMSE) and the (SMSE) for the proposed generalized biased estimator, we use the spectral decomposition of estimated information matrix $(X'\widehat{W}X)$.

Suppose that there exists a matrix ψ such that: $\psi(X'\widehat{W}X)\psi' = \Lambda = diag\{\lambda_j\}$, where $\lambda_1 \geq \lambda_2 \geq .. \geq \lambda_{p^*}$ are the ordered eigenvalues of $(X'\widehat{W}X)$ and ψ is a $p^* \times p^*$ orthogonal matrix whose columns are the corresponding eigenvectors of $\lambda_1, \lambda_2, \dots, \lambda_{p^*}$. The (MMSE) and the (SMSE) of $\hat{\beta}_{MLE}$ are as follow:

$$MMSE(\hat{\beta}_{MLE}) = \frac{1}{\phi}(X'\widehat{W}X)^{-1} = \frac{1}{\phi}\psi\Lambda^{-1}\psi' \quad (19)$$

$$SMSE(\hat{\beta}_{MLE}) = tr(Cov(\hat{\beta}_{MLE})) = \frac{1}{\phi}\sum_{j=1}^{p*}\frac{1}{\lambda_j} \quad (20)$$

Let $\beta = \psi\alpha,\ \alpha = \psi'\beta$, so that the proposed generalized biased estimator can be defined as:

$$\hat{\alpha}_{GBRRE} = (\Lambda + K)^{-1}(\Lambda)\hat{\alpha}_{MLE} \quad (21)$$

Note that, $\hat{\beta}_{MLE} = \psi'\hat{\alpha}_{MLE}$. Let $\Lambda_K = (\Lambda + K)$, The matrix MSE for $\hat{\beta}_{GBRRE}$ is:

$$MMSE(\hat{\beta}_{GBRRE}) = Cov(\hat{\beta}_{GBRRE}) + bias(\hat{\beta}_{GBRRE})bias(\hat{\beta}_{GBRRE})'$$

$$MMSE(\hat{\beta}_{GBRRE}) =$$

$$\frac{1}{\phi}\psi\Lambda_K^{-1}\Lambda\Lambda_K^{-1}\psi' + bias(\hat{\beta}_{GBRRE})bias(\hat{\beta}_{GBRRE})' \quad (22)$$

$$bias(\hat{\beta}_{GBRRE}) = \psi\Lambda_K^{-1}\Lambda\psi'\beta - \beta = \psi\Lambda_K^{-1}\Lambda\alpha - \psi\alpha$$

$$bias(\hat{\beta}_{GBRRE}) = \psi[\Lambda_K^{-1}\Lambda - I]\alpha \quad (23)$$

$$bias(\hat{\beta}_{GBRRE})bias(\hat{\beta}_{GBRRE})' =$$

$$= \psi[\Lambda_K^{-1}\Lambda - I]\alpha\alpha'[\Lambda_K^{-1}\Lambda - I]' \quad (24)$$

The scalar MSE for $\hat{\beta}_{GBRRE}$ is:

$$SMSE(\hat{\beta}_{GBRRE}) =$$

$$tr(Cov(\hat{\beta}_{GBRRE})) + bias(\hat{\beta}_{GBRRE})'bias(\hat{\beta}_{GBRRE}) \quad (25)$$

$$SMSE(\hat{\beta}_{GBRRE}) = \frac{1}{\phi}\sum_{j=1}^{p*}\frac{\lambda_j}{(\lambda_j+k_j)^2} + \sum_{j=1}^{p*}\frac{\alpha_j^2k_j^2}{(\lambda_j+k_j)^2} \quad (26)$$

2.4. Comparison between $\hat{\beta}_{\text{MLE}}$ and $\hat{\beta}_{\text{GBRRE}}$

Let $\hat{\theta}_1, \hat{\theta}_2$ be two estimators of θ, the estimator $\hat{\theta}_1$ is said to be superior to the estimator $\hat{\theta}_2$ in the sense of MMSE criterion, if and only if:

$$\Delta(\hat{\theta}_1, \hat{\theta}_2) = MMSE(\hat{\theta}_1) - MMSE(\hat{\theta}_2) \geq 0$$

Lemma 2: Let A be a positive definite matrix, τ be a vector of nonzero constants and c be a positive constant, then $cA - \tau\tau' > 0$ if and only if $\tau' A\, \tau < \text{c}$ (Farebrother, 1976).

Lemma3: Let $\hat{\theta}_1, \hat{\theta}_2$ be two estimators of θ, $\hat{\theta}_1 = B_1y$, and $\hat{\theta}_2 = B_2y$. Assume that $\Delta = cov(\hat{\theta}_1) - cov(\hat{\theta}_2) > 0$, where $cov(\hat{\theta}_1), cov(\hat{\theta}_2)$ represent the covariance matrices of the estimators $\hat{\theta}_1, \hat{\theta}_2$, respectively. Then, $\Delta(\hat{\theta}_1, \hat{\theta}_2) = MMSE(\hat{\theta}_1) - MMSE(\hat{\theta}_2) \geq 0$, if and only if:

$b_2'(\Delta + b_1b_1')^{-1}b_2 \leq 1$, where b_1, b_2 denote the bias vectors of the estimators $\hat{\theta}_1, \hat{\theta}_2$, respectively see Farebrother [24], Trenkler and Toutenburg [38].

Theorem1: Let $K = diag\{k_j\}$; $k_j > 0,\ j = 1,2, \dots, p^*$, then $\hat{\beta}_{\text{GBRRE}}$ is superior to the estimator $\hat{\beta}_{\text{MLE}}$ in the sense of MMSE criterion, i.e. $MMSE(\hat{\beta}_{\text{MLE}}) - MMSE(\hat{\beta}_{\text{GBRRE}}) > 0$ if

$$bias(\hat{\beta}_{GBRRE})'[\Lambda^{-1} - \Lambda_K^{-1}\Lambda\Lambda_K^{-1}]bias(\hat{\beta}_{GBRRE}) < 1.$$

Proof: the difference between $MMSE(\hat{\beta}_{\text{MLE}})$ *and* $MMSE(\hat{\beta}_{\text{GBRRE}})$ is calculated as follows:

$$MMSE(\hat{\beta}_{\text{MLE}}) - MMSE(\hat{\beta}_{\text{GBRRE}}) = \frac{1}{\phi}\psi(\Lambda^{-1} - \Lambda_K^{-1}\Lambda\Lambda_K^{-1})\psi' + $$

$$+ bias(\hat{\beta}_{GBRRE})bias(\hat{\beta}_{GBRRE})' \quad (27)$$

$$MMSE(\hat{\beta}_{\text{MLE}}) - MMSE(\hat{\beta}_{\text{GBRRE}}) =$$

$$= \frac{1}{\phi}\psi\ diag\ \left\{\frac{1}{\lambda_j} - \frac{\lambda_j}{(\lambda_j+k_j)^2}\right\}_{j=1}^{p*}\psi' \quad (28)$$

Since $bias(\hat{\beta}_{GBRRE})bias(\hat{\beta}_{GBRRE})'$ in Eq (27) is nonnegative definite. Then, the matrix

$(\Lambda^{-1} - \Lambda_K^{-1}\Lambda\Lambda_K^{-1})$ is positive definite if $(\lambda_j + k_j)^2 > \lambda_j$, i.e. $\lambda_j^2 + k_j^2 + \lambda_j(2k_j - 1) > 0$, thus for $k_j > 0, j = 1,2, \dots, p^*$ the proof is completed by lemma 3. Hence, it can be concluded that when the multicollinearity problem exists then the proposed generalized biased estimator $(\hat{\beta}_{GBRRE})$ is better than $(\hat{\beta}_{MLE})$ for the beta regression model because it has a smaller covariance matrix.

3. Selection of Shrinkage Parameters (k_j)

The idea of a generalized biased estimator is to choose proper values of k_j such that the decrease in the covariance terms is greater than the increase of the squared bias. Thus, $SMSE(\hat{\beta}_{GBRRE})$ and will be less than $SMSE(\hat{\beta}_{MLE})$. Eq (26) can be written be as follows:

$$SMSE(\hat{\beta}_{GBRRE}) = \xi_1(k_j) + \xi_2(k_j) \quad (29)$$

The first term $\xi_1(k_j)$ is the covariance and the second term $\xi_2(k_j)$ is the squared bias. It is evident that the functions in equation (29) are continuous functions of k_j, so we can obtain the optimal values of k_j that minimizes $SMSE(\hat{\beta}_{GBRRE})$ by differentiating Eq (26) with respect to d_j as follows:

$$\frac{\partial\, SMSE(\hat{\beta}_{GBRRE})}{\partial k_j} = \frac{-2}{\phi}\sum_{j=1}^{p*}\frac{\lambda_j}{(\lambda_j+k_j)^3} + 2\sum_{j=1}^{p*}\frac{k_j\alpha_j^2\lambda_j}{(\lambda_j+k_j)^3}$$

$$\frac{\partial\, SMSE(\hat{\beta}_{GBRRE})}{\partial k_j} = \frac{2}{\phi}\left[\sum_{j=1}^{p*}\frac{\alpha_j^2 k_j\lambda_j\phi}{(\lambda_j+k_j)^3} - \sum_{j=1}^{p*}\frac{\lambda_j}{(\lambda_j+k_j)^3}\right]$$

Then equating the derivative to zero, with simplification,

$$\therefore \quad k_{j(optimal)} = \frac{1}{\alpha_j^2\phi}\ , \qquad j = 1,2,\dots,p^* \qquad (30)$$

We suggest the following new estimation of k_j:

$$k_{j(1)} = \frac{\phi}{\hat{\alpha}_j^2}\ ,\quad k_{j(2)} = \frac{\lambda_{max}}{\hat{\alpha}_j^2\phi}\ ,\quad k_{j(3)} = \frac{\lambda_{min}}{\hat{\alpha}_j^2\phi}$$

where $\hat{\alpha}_j = \psi'\hat{\beta}_{j(MLE)}$ and λ_{max} and λ_{min} are the maximum and the minimum eigenvalues of $(X'\widehat{W}X)$.

4. Simulation

In order to compare the performance of different estimators, a simulation study is introduced in this section.

4.1. The Design of the Experiment

The dependent variable is generated using the following formula:

$$Y_i \sim B[\mu(X_i), \varphi] \qquad (31)$$

where φ is the precision parameter that controls variability; in this simulation study the precision parameter used is: $\varphi = 0.5, 1, 1.5$.

The number of explanatory variables used in this study is: p = 2, 4, 6. the explanatory variables were generated using the following equation see Abonazel [3]:

$$x_{ij} = (1-\rho^2)^{1/2} z_{ij} + \rho z_{ip} \qquad (32)$$

Where, $i = 1,\dots,n$; $j = 1,\dots,p$ and z_{ij} are generated using the standard normal distribution, and ρ is specified so that the correlation between any two explanatory variables is given by ρ^2. Three different values of ρ^2 are considered corresponding to

$\rho = 0.85, 0.90,$ and 0.95 .The parameter values of β are chosen so that:

$$\acute{\beta}\beta = 1; \quad \beta_1 = \beta_2 = \cdots = \beta_P.$$

the sample size used is: n= 20, 50, 100. The total number of Monte Carlo replications is R=1000.

The R programming language, based on betareg package Cribari-Neto and Zeileis [20], is used to conduct our simulation study.

4.2. Performance Criteria

The comparison is based on $MSE(\beta)$ which is calculated as follows:

$$MSE(\beta) = \frac{\sum_{i=1}^{R}(\hat{\beta}-\beta)'(\hat{\beta}-\beta)}{R}$$

where $\hat{\beta}$ is the estimator of β obtained from MLE, ridge, and different K's.

4.3. Result and Discussion

Tables 1, 2, 3 contain the results for sample sizes n = 20, 50, 100. We note that:

1. The MSE of beta regression estimators are smaller than that of ML estimator
2. In general, as the correlation coefficient increases, the $MSE(\beta)$ is increasing for all estimators.
3. By increasing the number of explanatory variables, the MSE values of all the estimation methods increases.
4. All the proposed methods of k outperform the ML estimator in terms of MSE, especially k1
5. The best estimator is k1 for all values of ρ; and the worst one is K2.

Table 1. MSE when n =20

φ	p	ρ	MLE	Ridge	K1	K2	K3
0.5	2	0.85	0.151078	0.130115	0.0849323	1.094954	0.085779
		0.90	0.889005	0.439220	0.1004256	1.142654	0.1232848
		0.95	8.135810	5.398778	2.6449510	13.51673	1.3099693
	4	0.85	0.813820	0.233127	0.1421025	0.7451736	0.1787042
		0.90	12.49408	10.00071	1.781790	16.90343	5.497808
		0.95	17.55270	15.88778	6.685765	20.3331	6.955368
	6	0.85	1.191371	1.007452	0.2785795	1.424297	0.332215
		0.90	6.075889	5.514079	4.667240	6.413113	4.675621
		0.95	17.52226	15.69885	10.15286	16.67643	10.389902
1	2	0.85	0.577835	0.210078	0.1584821	0.531461	0.1969791
		0.90	0.641091	0.546352	0.5819947	0.7946418	0.5847204
		0.95	11.24824	5.250160	11.03252	14.56423	11.4231651
	4	0.85	0.695462	0.355016	0.105190	0.6927259	0.1334194
		0.90	5.182378	4.016946	0.7469011	4.671093	3.59651
		0.95	23.36121	18.26985	4.436750	19.75574	20.96318
	6	0.85	1.009363	0.690398	0.1318896	0.8755702	0.2345632
		0.90	2.874960	1.941652	0.228058	2.272636	0.9980387
		0.95	7.832783	4.618213	1.334263	7.122937	4.143187
1.5	2	0.85	0.279418	0.261176	0.06109178	0.6630706	0.1664093
		0.90	0.799001	0.300570	0.1396927	0.9347021	0.2956866
		0.95	10.80926	8.320316	2.467042	8.580005	7.9744
	4	0.85	8.059327	3.766810	0.560573	6.222027	2.971153
		0.90	8.901262	6.966372	3.148902	7.372183	6.535305
		0.95	14.17223	8.788175	6.746116	13.63090	6.780772
	6	0.85	3.672061	2.500191	0.7166938	4.229359	0.852210
		0.90	6.102034	3.133449	0.907870	4.723997	2.835343
		0.95	6.357930	5.400654	1.825705	5.436429	5.300449

Table 2. MSE when n =50

φ	p	ρ	MLE	Ridge	K1	K2	K3
0.5	2	0.85	0.261545	0.111595	0.07018114	1.172829	0.1844499
		0.90	0.270571	0.172406	0.1627681	1.172907	0.2017397
		0.95	0.515019	0.3640839	0.1870535	1.249902	0.2329233
	4	0.85	1.001133	0.8282114	0.06001644	1.262104	0.2030102
		0.90	1.395321	1.263944	0.2703892	1.800976	0.2889573
		0.95	2.490422	1.820516	0.7361842	2.264177	1.091452
	6	0.85	1.500124	1.425059	0.1866986	1.598779	0.7489884
		0.90	2.040702	1.856892	1.016292	2.13744	1.350609
		0.95	9.255347	7.975472	3.848493	8.677811	4.530881
1	2	0.85	0.191408	0.0908735	0.03383087	0.9396146	0.05989122
		0.90	0.496433	0.123962	0.07957709	0.9879783	0.1413697
		0.95	1.196744	0.7190675	0.6242875	2.194484	0.7871522
	4	0.85	0.160914	0.1106989	0.1230966	0.5172635	0.1366009
		0.90	0.685891	0.4755008	0.1736001	0.8138175	0.3174097
		0.95	4.23645	2.925781	0.2819842	5.047632	2.193098
	6	0.85	0.209517	0.1380256	0.03260192	0.5101675	0.04580345
		0.90	0.575926	0.3852131	0.1310574	0.6687215	0.1362897
		0.95	1.413364	0.8805163	0.9168111	1.313029	1.089953
1.5	2	0.85	0.364643	0.2057379	0.1788385	0.8052479	0.1801431
		0.90	0.406281	0.1938494	0.1474229	0.7966347	0.1620081
		0.95	1.636993	0.3842578	0.1965658	1.10943	0.5725826
	4	0.85	0.195896	0.0669939	0.06176634	0.4545175	0.08107146
		0.90	0.603136	0.4298793	0.1179199	0.6092083	0.1231678
		0.95	5.058025	3.1134720	0.1578872	5.051569	3.17643
	6	0.85	0.164371	0.07410614	0.07675847	0.3818574	0.08730359
		0.90	0.804897	0.4857059	0.174093	0.7174643	0.1970021
		0.95	5.507328	4.399793	3.923305	5.248523	4.2897480

Table 3. MSE when n =100

φ	p	ρ	MLE	Ridge	K1	K2	K3
0.5	2	0.85	0.2658368	0.2034754	0.07006755	1.092735	0.1013392
		0.90	0.4702379	0.2931466	0.1085491	1.257625	0.184965
		0.95	0.6106628	0.4008274	0.2747804	1.601246	0.3653858
	4	0.85	0.2686006	0.206739	0.03998615	0.8447301	0.2028978
		0.90	0.4998536	0.4606279	0.1254036	1.038591	0.3574079
		0.95	0.7590143	0.4631254	0.2044859	1.153067	0.5393513
	6	0.85	0.3531833	0.3179042	0.115634	0.6232421	0.2443423
		0.90	0.8395655	0.7639008	0.2737073	1.00460	0.4379536
		0.95	2.252314	2.007729	0.5068555	2.676599	0.6553334
1	2	0.85	0.3673951	0.1889065	0.05433793	1.00416	0.07026247
		0.90	0.5055386	0.3942442	0.1533143	1.194422	0.3255137
		0.95	3.117502	1.936752	0.4498242	5.704797	1.3318000
	4	0.85	0.185144	0.1337849	0.06795854	0.5793321	0.0966373
		0.90	0.4730394	0.3573018	0.08610637	0.6819047	0.1081919
		0.95	1.269237	0.7479217	0.1548721	1.362956	0.2846773
	6	0.85	0.1060956	0.07409421	0.04776824	0.4285647	0.0729202
		0.90	0.566404	0.4290715	0.2590651	0.8749183	0.3003917
		0.95	1.686679	1.231194	0.3800645	2.215334	0.4517982
1.5	2	0.85	0.2120868	0.1236165	0.08544684	0.7451736	0.09554473
		0.90	0.2763646	0.1816971	0.1421025	0.789822	0.1787042
		0.95	0.8138205	0.2331266	0.176798	0.9006854	0.1892505
	4	0.85	0.2197505	0.1579647	0.1304972	0.4901639	0.161468
		0.90	0.6778576	0.6580156	0.5247445	1.09872	0.5338868
		0.95	5.348982	3.840387	0.6921845	4.100477	4.060253
	6	0.85	0.3911238	0.234933	0.0699931	0.6484935	0.1121890
		0.90	0.8472404	0.5686437	0.3282576	0.6861641	0.3341864
		0.95	1.048191	0.7436583	0.3682361	1.320537	0.8832337

5. Real Data Application

In this section, the beta regression model with different proposed estimators is applied to real data application. The beta regression model is applied to the Egyptian pre-university education during the academic year 2018/2019. The real data was obtained from the annual bulletin for pre-university education, which published by the Egyptian central agency for public mobilization and statistics (CAPMAS), November 2019.

The dependent variable Y is the proportion of pre-university students to the total population of the governorate (we have 27 governorates in Egypt), this proportion of students affected by six explanatory variables which are: the number of schools in the governorate (X_1); the number of classes (X_2); the number of male students (X_3); the number of female students (X_4); the number of male teachers (X_5); and the number of female teachers (X_6).

Table 4 shows some descriptive measures for the above dependent and explanatory variables.

Table 4. Descriptive statistics; VIF; and eigen values for the variables

Variable	Symbol	Mean	Standard deviation	VIF	Eigen values
No. of schools	X_1	135.48	144.66	95.96	85.19123
No. of classes	X_2	1605.19	1582.113	566.54	1.287482
No. of male students	X_3	29506.15	30050.23	583.64	1.181225
No. of female students	X_4	35453.56	33898.68	349.42	0.357269
No. of male teachers	X_5	2262.59	1799.33	25.87	0.026723
No. of female teachers	X_6	1703.11	1686.69	25.342	0.014333
proportion of students	Y	0.037111	3.63	—	—

The above variables are highly correlated according to the following correlation matrix:

$$\begin{bmatrix} 1 & .986 & .974 & .960 & .923 & .944 \\ .986 & 1 & .994 & .989 & .963 & .960 \\ .974 & .994 & 1 & .996 & .958 & .936 \\ .960 & .989 & .996 & 1 & .968 & .939 \\ .923 & .963 & .958 & .968 & 1 & .933 \\ .944 & .960 & .936 & .939 & .933 & 1 \end{bmatrix}$$

Figure 1 shows the Correlogram of the above correlation matrix. These circles show both the sign and magnitude of the correlation value. Color is used to encode the sign of the correlation (blue for positive values, red for negative values).

These circles fill an area proportional to the absolute value of the correlation. The circles are filled clockwise for positive values, anti-clockwise for negative values.

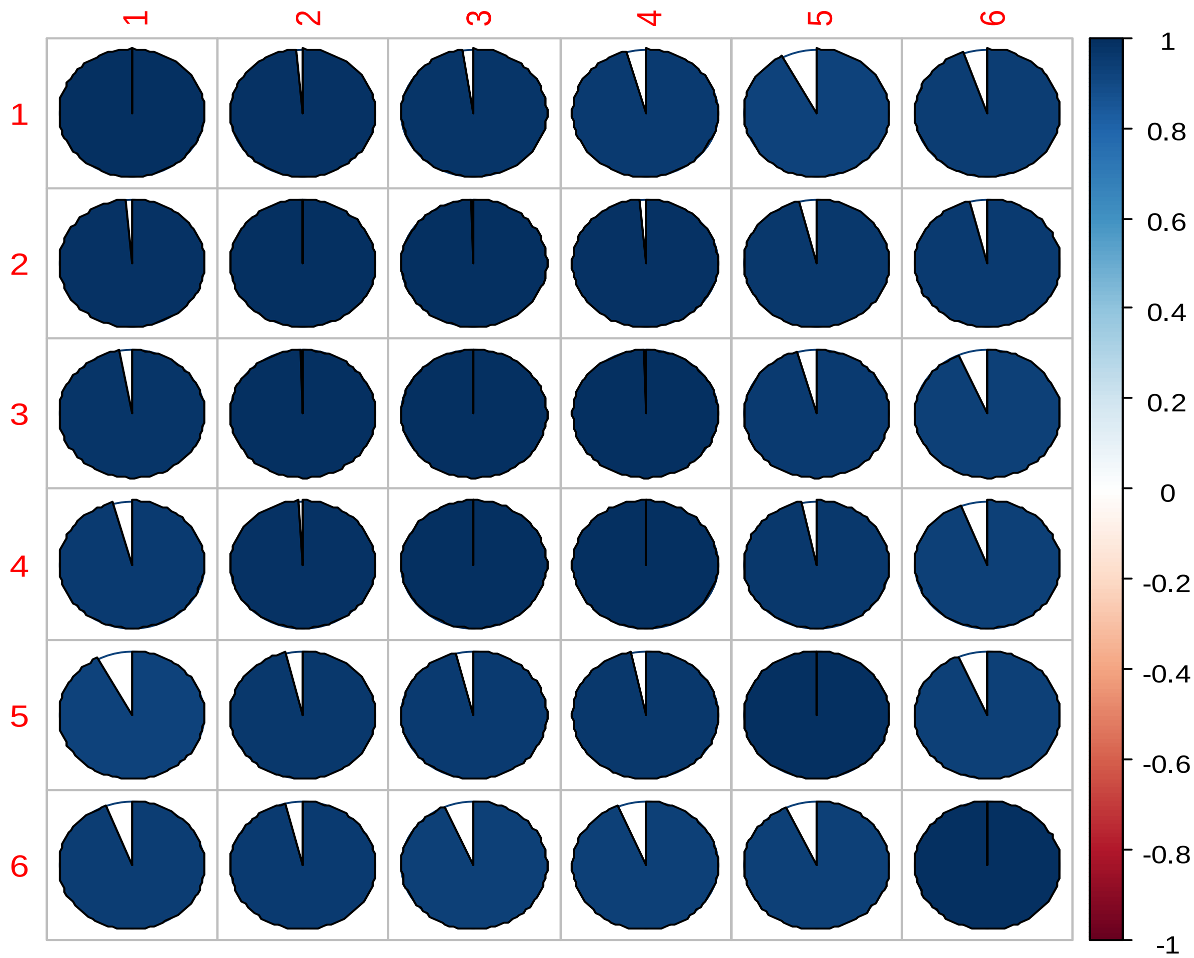

Figure 1. The correlogram for the correlation matrix

Table 5. Parameter estimates and estimated SMSE values of ML, Ridge, and different estimators

variable	MLE	Ridge	K1	K2	K3
X_1	-0.5099	-0.12512416	0.5427595	-0.200159	0.541440
X_2	0.7574	0.03678832	-0.749656	-0.685888	-0.755508
X_3	-0.5152	0.07567122	0.6230089	0.867269	0.536584
X_4	0.5689	0.14444665	-0.512541	-0.132328	-0.559546
X_5	0.1172	0.18075803	0.0688834	0.345954	-0.089369
X_6	-0.1075	0.00448434	0.3017812	0.1361004	0.1863447
SMSE	174.4148	1.295948	1.149797	4.218923	2.016722

Table 5 gives the parameter estimates for the beta regression model using ML method, ridge regression and with the different K's. The estimated SMSE were obtained using equations: (12) and (26).

From Table 5, we note that, according to the SMSE, the proposed K1 estimator outperform the ML estimators and ridge estimators, and other K's.

6. Conclusions

In this paper, the multicollinearity problem in beta regression model is investigated. We propose a generalized ridge regression estimator for correcting multicollinearity. The performance of the proposed generalized biased estimator is evaluated theoretically and practically. According to the simulation study, we found that the optimal estimator is K1 and the worst one is K2. Finally, we apply the different estimator to the pre-university education data in Egypt during the academic year (2018/2019), and also, we found that the optimal estimator is K1.

Acknowledgement

We are very grateful to reviewers for their appropriate and constructive suggestions to improve this paper.

REFERENCES

[1] Abonazel M., Farghali R. Liu-type multinomial logistic regression. Sankhya B 81, pp. 203–225, 2019. DOI: https://doi.org/10.1007/s13571-018-0171-4

[2] Abonazel M., Taha I. Beta ridge regression estimators: Simulation and Application. Communications in Statistics-Simulation and Computation, 2021. DOI: https://doi.org/10.1080/03610918.2021.1960373.

[3] Abonazel, M., A practical guide for creating Monte-Carlo simulation studies using R. International Journal of Mathematics and Computational Science 4, pp.18–33, 2018.

[4] Akram M., Amin M., Elhassanein A., and Aman Ullah M., A new modified ridge-type estimator for the beta regression model: simulation and application. AIMS Mathematics 7, pp.1035–1057, 2021a. DOI: https://doi.org/10.3934/math.2022062

[5] Akram M., Amin M., Faisal M., On the generalized biased estimators for the gamma regression model: methods and applications, Communications in Statistics- Simulation and Computation, 2021b. DOI: https://doi.org/10.1080/03610918.2021.1953073.

[6] Akram, M. N., Amin, M., & Qasim, M., A new Liu-type estimator for the Inverse Gaussian regression model. Journal of Statistical Computation and Simulation, pp. 1-20, 2020. DOI: https://doi.org/10.1080/00949655.2020.1718150

[7] Algamal Z. Developing a ridge estimator for the gamma regression model, Journal of Chemometrics 32: e3054, 2018. DOI: https://doi.org/10.1002/cem.3054

[8] Algamal Z., Abonazel M. Developing a Liu-type estimator in beta regression model, Concurrency and Computation-Practice and Experience, 2021. DOI: https://doi.org/10.1002/cpe.6685.

[9] Alheety M., Kibria G. On the Liu and almost unbiased Liu estimators in the presence of multicollinearity with heteroscedastic or correlated errors. Surveys in Mathematics and its Applications 4, pp. 155 – 167, 2009.

[10] Amin M., Akram M., and Amanullah M. On the James-Stein estimator for the Poisson regression model. Communications in Statistics: Simulation and Computation, 2021a. DOI:https://doi.org/10.1080/03610918.2020.177585

[11] Amin M., Nauman M., and Kibria G. A new adjusted Liu estimator for the Poisson regression model, Concurrency and Computation- Practice and Experience, 2021B, DOI: https://doi.org/10.1002/cpe.6340.

[12] Amin, M., Qasim, M., and Amanullah, M. Performance of Asar and Genç and Huang and Yang's Two-Parameter Estimation Methods for the Gamma Regression Model. Iranian Journal of Science and Technology, Transactions A: Science, 43, pp. 2951-2963, 2019. DOI: https://doi.org/10.1007/s40995-019-00777-3

[13] Amin, M., Qasim, M., Amanullah, M., & Afzal, S. Performance of some ridge estimators for the gamma

regression model. Statistical papers 61, pp. 997-1026. DOI: https://doi.org/doi.org/10.1007/s00362-017-0971-z

[14] Asar Y. Liu-type logistic estimators with optimal shrinkage parameter. Journal of Modern Applied Statistical Methods 15, pp.738-751, 2016. DOI: https://doi.org/10.22237/jmasm/1462077300

[15] Asar Y., and Genç A., New shrinkage parameters for the Liu-type logistic estimators. Communications in Statistics-Simulation and Computation 45, pp.1094-1103, 2016. DOI: https://doi.org/10.1080/03610918.2014.995815

[16] Asar, Y., & Genç, A., A new two-parameter estimator for the poisson regression model. Iranian Journal of Science and Technology, Transactions A: Science, 42(2), pp.793-803, 2018. DOI: https://doi.org/10.1007/s40995-017-0174-4

[17] Awwad F., Dawoud I., and Abonazel M., (2021), Development of robust Özkale–Kaçiranlar and Yang–Chang estimators for regression models in the presence of multicollinearity and outliers, Concurrency and Computation- Practice and Experience, DOI: https://doi.org/10.1002/cpe.6779.

[18] Belsley D. A guide to using the collinearity diagnostics. Computer Science in Economics and Management 4, p. 33-50, 1991.

[19] Bhat R. and Vidya S. A class of generalized ridge estimator. Communications in Statistics-Simulation and Computation 47, pp.1094-1103, 2017. DOI: https://doi.org/10.1080/03610918.2016.1144765

[20] Cribari-Neto F., and Zeileis A. Beta regression in R. Journal of Statistical Software 34 pp. 1–24, 2010. DOI: https://doi.org/10.18637/jss.v034.i02

[21] Eledum H., Awadallah H. A Mont Carlo study for dealing with multicollinearity and autocorrelation problems in linear regression using two stage ridge regression methods. Mathematics and Statistics 9, pp. 630 – 638. DOI: https://doi.org/10.13189/ms.2021.090502

[22] Espinheira P., Ferrari S., Cribari–Neto F. Influence diagnostics in beta regression. Computational Statistics and Data Analysis 52, pp. 4417–4431, 2008b. DOI: https://doi.org/10.1016/j.csda.2008.02.028

[23] Espinheira, P., Ferrari S., and Cribari–Neto. On beta regression residuals. Journal of Applied Statistics 35, pp. 407– 419, 2008a. DOI: https://doi.org/10.1080/02664760701834931

[24] Farebrother, R. W. Further results on the mean square error of ridge regression. Journal of the Royal Statistical Society. Series B (Methodological) 38, pp. 248-250, 1976.

[25] Farghali R., Qasim M., Kibria G., and Abonazel M., Generalized two-parameter estimators in the multinomial logit regression model: methods, simulation and application. Communications in Statistics- Simulation and Computation, 2021. DOI: https://doi.org/10.1080/03610918.2021.1934023.

[26] Farghali, R. "Generalized Liu-Type estimator for linear regression". International Journal of Research and Reviews and in Applied Sciences 38, pp. 52-63, 2019.

[27] Ferrari S., and Cribari-Neto F. "Beta regression for modeling rates and proportions". Journal of Applied Statistics 31, pp. 799-815, 2004. DOI: https://doi.org/10.1080/0266476042000214501

[28] Hoerl A., and Kennard W. "Ridge regression: biased estimation for nonorthogonal problems". Technometrics 12, pp.55-67, 1970a DOI: https://doi.org/10.1080/00401706.1970.10488634.

[29] Hoerl A., and Kennard W. "Ridge regression: applications to nonorthogonal problems". Technometrics 12, pp.69-82, 1970b. DOI: https://doi.org/10.2307/1267352.

[30] Issa M. "Weighted least squares estimation for AR(1) model with incomplete data", Mathematics and Statistics 10, pp. 342 – 357. DOI: https://doi.org/10.13189/ms.2022.100209

[31] Karlsson P., Månsson K., and Kibria G. "A Liu estimator for the beta regression model and its application to chemical data", Journal of Chemometrics 34: e3300, 2020. DOI: https://doi.org/10.1002/cem.3300.

[32] Kieschnick R., and McCullough D. "Regression analysis of variates observed on (0, 1): percentages, proportions and fractions", Statistical Modelling 3, p.193–213, 2003. DOI: https://doi.org/10.1191/1471082X03st053oa

[33] Liu K. "A new class of biased estimate in linear regression". Communications in Statistics-Theory and Methods 22, p.393-402, 1993. DOI: https://doi.org/10.1080/03610929308831027

[34] Liu K. "Using Liu-type estimator to combat collinearity". Communications in Statistics-Theory and Methods 32, p.1009-2003, 2003. https://doi.org/10.1081/STA-120019959.

[35] Liu K.. "More on Liu-type estimator in linear regression". Communications in Statistics-Theory and Methods 33, pp.2723-2733, 2004. https://doi.org/10.1081/STA-200037930

[36] Lukman A., Adewuyi E., Månsson K., and Kibria G. "A new estimator for the multicollinear Poisson regression model: simulation and application", Scientific Reports 11, 3732, 2021. https://doi.org/10.1038/s41598-021-82582-w.

[37] Qasim M., Månsson K., and Kibria G. "On some beta ridge regression estimators: method, simulation and application". Journal of Statistical Computation and Simulation 91, pp. 1699-1712, 2021. DOI: https://doi.org/10.1080/00949655.2020.1867549

[38] Trenkler G., and Toutenburg H. "Mean squared error matrix comparisons between biased estimators—An overview of recent results". Statistical Papers 31, pp. 165, 1990. DOI: https://doi.org/10.1007/BF02924687

[39] Ünlüa H., and Aktaş S. "Beta regression for the indicator values of well-being index for provinces in Turkey". Journal of Engineering Technology and Applied Sciences 2, pp. 101-111, 2017. DOI: https://doi.org/10.30931/jetas.321165