THE PHILIPPINE STATISTICIAN

An Official Publication of the
Philippine Statistical Association Inc.

Volume 70, Number 2 (2021)

Indexed in Scopus since 2015

The Philippine Statistician is the official scientific journal of the Philippine Statistical Association. Inc. (PSAI). It considers papers resulting from original research in statistics and its applications. Papers will be sent for review on the assumption that this has not been published elsewhere nor is submitted in another journal.

ISSN 2094-0343

THE PHILIPPINE STATISTICIAN

Volume 70, Number 2 (2021)

The Official Publication of the
Philippine Statistical Association, Inc.

Indexed in Scopus since 2015

Contents

Editorial

The second publication of the 70th volume of *The Philippine Statistician* includes five papers exploring applications of statistical theories and methods. Z. Algamal and co-authors propose a modified logistic ridge estimator to decrease shrinkage parameter and improve the resultant estimator with small bias. S. Dar and co-authors illustrate a new compound probability model applicable by compounding Poisson distribution with two parameter Pranav distribution to count data while S. Khare and co-authors analyze classes of estimators under new calibration schemes using non-conventional measures of dispersion. Z. F. Althobaiti and A. Shabri investigate the economic aspects of gas emissions and predict CO_2 emissions using annual time series data in Saudi Arabia. M. Ghalibaf presents two new tests for tail independence in extreme value models.

This publication will not be possible without the time, effort and expertise of our editorial board members, the editorial staff, the secretariat and anonymous reviewers. My gratitude also go to the authors of the papers in this journal, as well as other authors of papers that have undergone review for publication. To the authors of the papers who have successfully gone through the editorial process, the editorial staff of the journal highly appreciate your contributions to push research in Statistics to greater heights. Everyone's contributions help in preserving the quality and integrity of the publication. Our journal editors will continue to uphold the level of trust bestowed to *The Philippine Statistician* for its quality.

Jose Ramon G. Albert
Editor-in-Chief

A Modified Ridge Estimator for the Logistic Regression Model

Mazin M. Alanaz
Department of Operation Research and Intelligence Techniques, University of Mosul, Iraq.

Nada Nazar Alobaidi
Department of Statistics and Informatics, University of Mosul, Mosul, Iraq

Zakariya Yahya Algamal*
Department of Statistics and Informatics University of Mosul, Iraq

The ridge estimator has been consistently demonstrated to be an attractive shrinkage method to reduce the effects of multicollinearity. The logistic regression model is a well-known model in application when the response variable is binary data. However, it is known that multicollinearity negatively affects the variance of maximum likelihood estimator of the logistic regression coefficients. To address this problem, a logistic ridge regression model has been proposed by numerous researchers. In this paper, a modified logistic ridge estimator (MLRE) is proposed and derived. The idea behind the MLRE is to get diagonal matrix with small values of diagonal elements that leading to decrease the shrinkage parameter and, therefore, the resultant estimator can be better with small amount of bias. Our Monte Carlo simulation results suggest that the MLRE estimator can bring significant improvement relative to other existing estimators.

Keywords: *multicollinearity, ridge estimator, logistic regression model, shrinkage, Monte Carlo simulation*

I. Introduction

Logistic regression model is widely applied for studying several real data problems, such as in medicine (Algamal and Lee 2015a). In dealing with the

* *Corresponding author: zakariya.algamal@uomosul.edu.iq*

logistic regression model, it is assumed that there is no correlation among the explanatory variables. In practice, however, this assumption often not holds, which leads to the problem of multicollinearity. In the presence of multicollinearity, when estimating the regression coefficients for logistic regression model using the maximum likelihood (ML) method, the estimated coefficients are usually become unstable with a high variance, and therefore low statistical significance (Kibria et al. 2015). Numerous remedial methods have been proposed to overcome the problem of multicollinearity. The ridge regression method (Hoerl and Kennard 1970) has been consistently demonstrated to be an attractive and alternative to the ML estimation method.

Ridge regression is a shrinkage method that shrinks all regression coefficients toward zero to reduce the large variance (Asar and Genç 2015; Rashad and Algamal 2019). This is done by adding a positive amount to the diagonal of $\boldsymbol{X}^T\boldsymbol{X}$. As a result, the ridge estimator is biased but it guaranties a smaller mean squared error than the ML estimator.

In linear regression, the ridge estimator is defined as

$$\hat{\boldsymbol{\beta}}_{Ridge} = (\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}, \tag{1}$$

where $\boldsymbol{y}$ is an n x 1 vector of observations of the response variable, $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_p)$ is an n x p known design matrix of explanatory variables, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$ is a p x 1 vector of unknown regression coefficients, $\mathbf{I}$ is the identity matrix with dimension p x p, and $k \geq 0$ represents the ridge parameter (shrinkage parameter). The ridge parameter, k, controls the shrinkage of $\boldsymbol{\beta}$ toward zero. The OLS estimator can be considered as a special estimator from Eq. (1) with $k = 0$. For larger value of k, the $\hat{\boldsymbol{\beta}}_{Ridge}$ estimator yields greater shrinkage approaching zero (Algamal and Lee 2015b; Hoerl and Kennard 1970).

2. Logistic Ridge Regression Model

Logistic regression is a statistical method to model a binary classification problem. The regression function has a nonlinear relation with the linear combination of the variables. In binary classification, the response variable of the logistic regression has two values either 1 for the tumor class, or 0 for the normal class. Let $\mathbf{y}_i \in \{0,1\}$ be a vector of size n x 1 of tissues, and let $\mathbf{x}_j$ be a p x 1 vector of variables. The logistic transformation of the vector of probability estimates $\pi_i = p(y_i = 1|\mathbf{x}_j)$ is modeled by a linear function, logit transformation,

$$\ln\left[\pi_i/1-\pi_i\right] = \beta_0 + \sum_{j=1}^{P} \mathbf{x}_j^T \beta_j, i = 1, 2, \dots, n, \tag{2}$$

where β_0 is the intercept, and β_j is a p x 1 vector of unknown variable coefficients. The log-likelihood function of Eq. (1) is defined as

$$\ell(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ \mathbf{y}_i \ln \pi(\mathbf{x}_{ij}) + (1 - \mathbf{y}_i) \ln(1 - \pi(\mathbf{x}_{ij}) \right\}. \tag{3}$$

Logistic regression offers the advantage of simultaneously estimating the probabilities $\pi(\mathbf{x}_{ij})$ and $1-\pi(\mathbf{x}_{ij})$ for each class and classifying subjects. The probability of classifying the i^{th} sample in class 1 is estimated by $\hat{\pi}_i = \exp\left(\beta_0 + \sum_{j=1}^{p} \mathbf{x}_j^T \beta_j\right) \Big/ 1 + \exp\left(\beta_0 + \sum_{j=1}^{p} \mathbf{x}_j^T \beta_j\right)$ (Algamal and Lee 2017; Algamal and Lee 2018; Algamal et al. 2017). The predicted class is then obtained by $I\{\hat{\pi}_i > 0.5\}$, where $I(\bullet)$ is an indicator function. The ML estimator is then obtained by computing the first derivative of the Eq. (2) and setting it equal to zero. Then, ML estimators of the logistic regression parameters (LRM) as

$$\hat{\boldsymbol{\beta}}_{LRM} = (\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{W}} \hat{\mathbf{v}}, \tag{4}$$

where $\hat{\mathbf{W}} = \mathrm{diag}(\hat{\theta}_i)$ and $\hat{\mathbf{v}}$ is a vector where i^{th} element equals to logit link function. The ML estimator is asymptotically normally distributed with a covariance matrix that corresponds to the inverse of the Hessian matrix

$$\mathrm{cov}(\hat{\boldsymbol{\beta}}_{LRM}) = \left[-E\left(\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_i \partial \beta_k} \right) \right]^{-1} = (\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X})^{-1}. \tag{5}$$

The mean squared error (MSE) of Eq. (5) can be obtained as

$$\begin{aligned} \mathrm{MSE}(\hat{\boldsymbol{\beta}}_{LRM}) &= E(\hat{\boldsymbol{\beta}}_{LRM} - \hat{\boldsymbol{\beta}})^T (\hat{\boldsymbol{\beta}}_{LRM} - \hat{\boldsymbol{\beta}}) \\ &= tr\left[(\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X})^{-1} \right] \\ &= \sum_{j=1}^{p} \frac{1}{\lambda_j}, \end{aligned} \tag{6}$$

where λ_j is the eigenvalue of the $\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X}$ matrix.

In the presence of multicollinearity, the matrix $\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X}$ becomes ill-conditioned leading to high variance and instability of the ML estimator of the Poisson regression parameters (Algamal 2018a; Algamal 2018b; Algamal and Alanaz 2018; Algamal and Asar 2018; Alkhateeb and Algamal 2020; Yahya Algamal 2018). As a remedy, Schaefer et al. (1984) proposed the logistic ridge regression model (LRRM) as

$$\hat{\boldsymbol{\beta}}_{LRRM} = (\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}\hat{\boldsymbol{\beta}}_{LRM}$$
$$= (\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^T\hat{\mathbf{W}}\hat{\mathbf{v}}, \quad (7)$$

where $k \geq 0$. The ML estimator can be considered as a special estimator from Eq. (7) with $k = 0$. Regardless of k value, the MSE of the $\hat{\boldsymbol{\beta}}_{LRRM}$ is smaller than that of $\hat{\boldsymbol{\beta}}_{LRM}$ because the MSE of $\hat{\boldsymbol{\beta}}_{LRRM}$ is equal to (Asar et al. 2017; Asar and Genç 2015; Kibria et al. 2012; Lukman et al. 2020; Månsson et al. 2011; Schaefer et al. 1984; Wu et al. 2016)

$$\text{MSE}(\hat{\boldsymbol{\beta}}_{LRRM}) = \sum_{j=1}^{p}\frac{\lambda_j}{(\lambda_j + k)^2} + k^2\sum_{j=1}^{p}\frac{\alpha_j}{(\lambda_j + k)^2}, \quad (8)$$

where α_j is defined as the j^{th} element of $\gamma\hat{\boldsymbol{\beta}}_{LRM}$ and γ is the eigenvector of the $\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}$ matrix. Comparing with the MSE of Eq. (6), $\text{MSE}(\hat{\boldsymbol{\beta}}_{LRRM})$ is always small for $k > 0$.

3. The New Estimator

In this section, the new estimator is introduced and derived. Let $\mathbf{M} = (m_1, m_2,\ldots, m_p)$ and $\Lambda = \text{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_p)$, respectively, "be the matrices of eigenvectors and eigenvalues of the $\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}$ matrix, such that $\mathbf{M}^T\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}\mathbf{M} = \mathbf{S}^T\hat{\mathbf{W}}\mathbf{S} = \Lambda$, where $\mathbf{S} = \mathbf{X}\mathbf{M}$. Consequently, the logistic regression estimator of Eq. (4), $\hat{\boldsymbol{\beta}}_{LRM}$, can be written as

$$\hat{\boldsymbol{\gamma}}_{LRM} = \Lambda^{-1}S^T\hat{\mathbf{W}}\hat{\mathbf{v}}$$
$$\hat{\boldsymbol{\beta}}_{LRM} = \mathbf{M}\hat{\boldsymbol{\gamma}}_{LRM}. \quad (9)$$

Accordingly, the logistic ridge estimator, $\hat{\boldsymbol{\beta}}_{LRRM}$, is rewritten as

$$\hat{\gamma}_{LRRM} = (\boldsymbol{\Lambda} + \mathbf{K})^{-1}\mathbf{S}^T\hat{\mathbf{W}}\mathbf{v}$$
$$= (\mathbf{I} - \mathbf{K}\mathbf{D}^{-1})\hat{\gamma}_{LRM}, \quad (10)$$

where $\mathbf{D} = \Lambda + \mathbf{K}$ and $\mathbf{K} = \text{diag}(k_1, k_2, \ldots, k_p)$; $k_i \geq 0$, $i = 1,2, \ldots, p$.

In generalized ridge estimator, the Jackknifing approach was used (Khurana et al. 2014; Nyquist 1988; Singh et al. 1986). Batah et al. (2008) proposed a modified Jackknifed ridge regression estimator in linear regression model.

In this paper, the modified estimator (MLRE) is derived by following the study of Batah et al. (2008). Let the Jackknife estimator (JE), in logistic regression, defined as

$$\hat{\gamma}_{JE} = (\mathbf{I} - \mathbf{K}^2\mathbf{D}^{-2})\hat{\gamma}_{LRM}, \tag{11}$$

and the modified Jackknife estimator (MJE) of Batah et al. (2008), in logistic regression model, is defined as

$$\hat{\gamma}_{MJE} = (\mathbf{I} - \mathbf{KD}^{-1})(\mathbf{I} - \mathbf{K}^2\mathbf{D}^{-2})\hat{\gamma}_{LRM}. \tag{12}$$

Consequently, our modified estimator is an improvement of Eq. (12) by multiplying it with the amount [($\mathbf{I}$–$\mathbf{K}^3\mathbf{D}^{-3}$) / ($\mathbf{I}$–$\mathbf{K}^2\mathbf{D}^{-2}$). The idea behind this is to get diagonal matrix with small values of diagonal elements which leading to decrease the shrinkage parameter, and, therefore, the resultant estimator can be better with small amount of bias. The new estimator is defined as

$$\hat{\gamma}MLRE = (\mathbf{I} - \mathbf{KD}^{-1})(\mathbf{I} - \mathbf{K}^2\mathbf{D}^{-2})\frac{(\mathbf{I} - \mathbf{K}^3\mathbf{D}^{-3})}{(\mathbf{I} - \mathbf{K}^2\mathbf{D}^{-2})}\hat{\gamma}LRM, \tag{13}$$

and

$$\hat{\boldsymbol{\beta}}_{MLRE} = \mathbf{M}^T\hat{\gamma}_{MLRE}. \tag{14}$$

4. Bias, Variance, and MSE of the New Estimator

The MSE of the new estimator can be obtained as

$$\text{MSE}(\hat{\gamma}_{MLRE}) = \text{var}(\hat{\gamma}_{MLRE}) + \left[\text{bias}(\hat{\gamma}_{MLRE})\right]^2 \tag{15}$$

According to Eq. (15), the bias and variance of $\hat{\gamma}_{MLRE}$ can be obtained as, respectively,

$$\begin{aligned}\text{bias}(\hat{\gamma}_{MLRE}) &= E\left[\hat{\gamma}_{MLRE}\right] - \gamma \\ &= (\mathbf{I} - \mathbf{KD}^{-1})(\mathrm{I} - \mathbf{K}^3\mathbf{D}^{-3})E\left[\hat{\gamma}_{MLRE}\right] - \gamma \\ &= -\mathbf{K}\left[(\mathbf{KD}^{-1})^{-1} - (\mathbf{KD}^{-1})^{-1}(\mathbf{I} - \mathbf{KD}^{-1}) + \mathbf{K}^2\mathbf{D}^{-2}(\mathbf{I} - \mathbf{KD}^{-1})\right]\mathbf{D}^{-1}\gamma,\end{aligned} \tag{16}$$

$$\begin{aligned}\text{var}(\hat{\gamma}_{MLRE}) &= (\mathbf{I} - \mathbf{KD}^{-1})(\mathbf{I} - \mathbf{K}^3\mathbf{D}^{-3})\text{var}(\hat{\gamma}_{MLRE})(\mathbf{I} - \mathbf{K}^3\mathbf{D}^{-3})^T(\mathbf{I} - \mathbf{KD}^{-1})^T \\ &= (\mathbf{I} - \mathbf{KD}^{-1})(\mathbf{I} - \mathbf{K}^3\mathbf{D}^{-3})\Lambda^{-1}(\mathbf{I} - \mathbf{K}^3\mathbf{D}^{-3})^T(\mathbf{I} - \mathbf{KD}^{-1})^T.\end{aligned} \tag{17}$$

Then,

$$\begin{aligned}\text{MSE}(\hat{\gamma}_{MLRE}) &= (\mathbf{I} - \mathbf{KD}^{-1})(\mathbf{I} - \mathbf{K}^3\mathbf{D}^{-3})\Lambda^{-1}(\mathbf{I} - \mathbf{K}^3\mathbf{D}^{-3})^T(\mathbf{I} - \mathbf{KD}^{-1})^T + \\ &\quad \left[-\mathbf{K}\left[(\mathbf{KD}^{-1})^{-1} - (\mathbf{KD}^{-1})^{-1}(\mathbf{I} - \mathbf{KD}^{-1}) + \mathbf{K}^2\mathbf{D}^{-2}(\mathbf{I} - \mathbf{KD}^{-1})\right]\mathbf{D}^{-1}\boldsymbol{\gamma}\right] \\ &\quad \left[-\mathbf{K}\left[(\mathbf{KD}^{-1})^{-1} - (\mathbf{KD}^{-1})^{-1}(\mathbf{I} - \mathbf{KD}^{-1}) + \mathbf{K}^2\mathbf{D}^{-2}(\mathbf{I} - \mathbf{KD}^{-1})\right]\mathbf{D}^{-1}\boldsymbol{\gamma}\right]^T \\ &= \boldsymbol{\Phi}\Lambda^{-1}\boldsymbol{\Phi}^T + \mathbf{K}\boldsymbol{\Psi}\mathbf{D}^{-1}\boldsymbol{\gamma}\boldsymbol{\gamma}^T\mathbf{D}^{-1}\boldsymbol{\Psi}^T\mathbf{K},\end{aligned} \tag{18}$$

where $\Phi = (\mathbf{I} - \mathbf{K}^3\mathbf{D}^{-3})^T (\mathbf{I} - \mathbf{K}\mathbf{D}^{-1})$ and $\Psi = [\mathbf{I} + \mathbf{K}\mathbf{D}^{-1} - \mathbf{K}\mathbf{D}^{-3}\mathbf{K}]$.

2.7. Selection of parameter k

The efficiency of ridge estimator strongly depends on appropriately choosing the *k* parameter. To estimate the values of *k* for our new estimator, the most well-known used estimation methods are employed and are given below (Kibria et al. 2015).

1. Hoerl and Kennard (1970) (HK), which is defined as

$$k_j(\text{HK}) = \frac{\hat{\sigma}^2}{\alpha_{\max}^2}, j = 1, 2, ..., p, \tag{19}$$

where $\hat{\sigma}^2 = \sum_{i=1}^{n} (y_i - \hat{\theta}_i)^2 / n - p - 1$.

2. Kibria et al. (2015) (KMS1), which is defined as

$$k_j(\text{KMSl}) = \text{Median}\left\{\left[\sqrt{\frac{\hat{\sigma}^2}{\hat{\alpha}_j^2}}\right]^2\right\}, j = 1, 2, ...p, \tag{20}$$

3. Kibria et al. (2015) (KMS2), which is defined as

$$k_j(\text{KMS2}) = \text{Median}\left\{\frac{\lambda_{\max}}{(n-p)\hat{\sigma}^2 + \lambda_{\max}\hat{\alpha}_j^2}\right\}, j = 1, 2, ...p, \tag{21}$$

5. Simulation Study

In this section, a Monte Carlo simulation experiment is used to examine the performance of the new estimator with different degrees of multicollinearity.

The response variable of *n* observations is generated from Bernoulli distribution regression model by

$$\pi_i = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})}, \tag{22}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ with $\sum_{j=1}^{p} \beta_j^2 = 1$ and $\beta_1 = \beta_2 = \ldots, = \beta_p$ (Kibria 2003; Månsson and Shukur 2011).

The explanatory variables $x_i^T = (x_{i1}, x_{i2}, \ldots, x_{in})$, have been generated from the following formula

$$x_{ij} = (1-\rho^2)^{1/2} w_{ij} + \rho w_{ip},\ i = 1, 2, \ldots, n,\ j = 1,2,\ldots, p, \tag{23}$$

where ρ represents the correlation between the explanatory variables and w_{ij}'s are independent standard normal pseudo-random numbers. Because the sample size has direct impact on the prediction accuracy, three representative values of the sample size are considered: 30, 50 and 100. In addition, the number of the explanatory variables is considered as $p=4$ and $p=8$ because increasing the number of explanatory variables can lead to increase the MSE. Further, because we are interested in the effect of multicollinearity, in which the degrees of correlation are considered more important, three values of the pairwise correlation are considered with $\rho = \{0.90,0.95,0.99)$. For a combination of these different values of n, p, and ρ, the generated data is repeated 1000 times and the averaged mean squared errors (MSE) is calculated as

$$\mathrm{MSE}(\hat{\boldsymbol\beta}) = \frac{1}{1000}\sum_{i=1}^{1000}(\hat{\boldsymbol\beta}-\boldsymbol\beta)^T(\hat{\boldsymbol\beta}-\boldsymbol\beta), \tag{23}$$

where $\hat{\boldsymbol\beta}$ is the estimated coefficients for the used estimator.

6. Simulation Results

The estimated MSE of Eq. (24) for MLE, LRM, and MLRE, for all the different selection methods of k and the combination of n, p, and ρ, are summarized in Tables 1, 2, and 3, respectively. Several observations can be made.

First, in terms of ρ values, there is increasing in the MSE values when the correlation degree increases regardless of the value of n, p. However, MLRE performs better than LRM and MLE for all the different selection methods of k. For instance, in Table 1, when $p = 8$ and $\rho = 0.99$, the MSE of MLRE was about 4.38%, 3.13%, and 2.86% lower than that of LRM for KH, KMS1 and KMS2, respectively. In addition, the MSE of MLRE was about 53.51% lower than that of MLE.

Second, regarding the number of explanatory variables, it is easily seen that there is increasing in the MSE values when the p increasing from four variables to eight variables. Although this increasing can affect the quality of an estimator, MLRE is achieved the lowest MSE comparing with MLE and LRM, for different n, p and different selection methods of k.

Third, with respect to the value of n, the MSE values decrease when n increases, regardless of the value of ρ, p, and the value of k. However, MLRE still consistently outperforms LRM and MLE by providing the lowest MSE.

Finally, for the different selection methods of *k,* the performance of all methods suggesting that the MLRE estimator is better than the other two estimators used. The KMS1 efficiently provides less MSE comparing with the KMS1 and KH for both MLRE and LRM estimators. Besides, KH is more efficient for providing less MSE than KMS2 or both MLRE and LRM estimators.

To summarize, all the considered values of *n*, *p*, *ρ*, and the value of *k*, MLRE is superior to LRM, clearly indicating that the new proposed estimator is more efficient.

Table 1. MSE values when *n* = 30

			KH		KMS1		KMS2	
	ρ	MLE	LRM	MLRE	LRM	MLRE	LRM	MLRE
p = 4	0.90	6.367	2.406	2.253	2.046	1.945	2.791	2.691
	0.95	6.995	2.637	2.486	2.495	2.394	2.952	2.849
	0.99	7.393	3.287	3.135	3.027	2.926	3.296	3.195
p = 8	0.90	6.472	2.608	2.455	2.238	2.137	2.986	2.885
	0.95	7.091	2.839	2.686	2.687	2.586	3.145	3.044
	0.99	7.506	3.489	3.336	3.219	3.118	3.491	3.391

Table 2. MSE values when *n* = 50

			KH		KMS1		KMS2	
	ρ	MLE	LRM	MLRE	LRM	MLRE	LRM	MLRE
p = 4	0.90	6.04	2.079	1.926	1.719	1.618	2.464	2.363
	0.95	6.668	2.312	2.159	2.168	2.067	2.623	2.522
	0.99	7.066	2.962	2.808	2.711	2.599	2.969	2.868
p = 8	0.90	6.145	2.281	2.128	1.911	1.811	2.659	2.558
	0.95	6.764	2.512	2.359	2.362	2.259	2.818	2.717
	0.99	7.179	3.162	3.009	2.892	2.791	3.164	3.063

Table 3. MSE values when *n* = 100

			KH		KMS1		KMS2	
	ρ	MLE	LRM	MLRE	LRM	MLRE	LRM	MLRE
p = 4	0.90	5.628	1.667	1.514	1.307	1.206	2.052	1.951
	0.95	6.256	1.898	1.747	1.756	1.655	2.211	2.112
	0.99	6.654	2.548	2.396	2.288	2.187	2.557	2.456
p = 8	0.90	5.733	1.869	1.716	1.499	1.398	2.247	2.146
	0.95	6.352	2.141	1.947	1.948	1.847	2.406	2.305
	0.99	6.767	2.751	2.597	2.481	2.379	2.752	2.651

7. Conclusion

In this paper, a modified estimator of logistic ridge regression is proposed to overcome the multicollinearity problem in the logistic regression model. According to Monte Carlo simulation studies, the modified estimator has a better performance than the maximum likelihood estimator and ordinary logistic ridge estimator, in terms of MSE. In conclusion, the use of the modified estimator is recommended when multicollinearity is present in the logistic regression model.

References

Algamal ZY. 2018a. "Developing a Ridge Estimator for the Gamma Regression Model." *Journal of Chemometrics* 32. doi:10.1002/cem.3054.

Algamal ZY. 2018b. "A New Method for Choosing the Biasing Parameter in Ridge Estimator for Generalized Linear Model." *Chemometrics and Intelligent Laboratory Systems* 183: 96-101.

Algamal ZY, Alanaz MM. 2018. "Proposed Methods in Estimating the Ridge Regression Parameter in Poisson Regression Model." *Electronic Journal of Applied Statistical Analysis* 11: 506-515.

Algamal ZY, Asar Y. 2018. "Liu-type Estimator for the Gamma Regression Model." *Communications in Statistics - Simulation and Computation* 49:2035-2048 doi:10.1080/03610918.2018.1510525.

Algamal ZY, Lee MH. 2015a. "High Dimensional Logistic Regression Model Using Adjusted Elastic Net Penalty." *Pakistan Journal of Statistics and Operation Research* 11: 667-676.

Algamal ZY, Lee MH. 2015b. "Penalized Poisson Regression Model Using Adaptive Modified." *Elastic Net Penalty Electronic Journal of Applied Statistical Analysis* 8: 236-245.

Algamal ZY, Lee MH. 2017. "A Novel Molecular Descriptor Selection Method in QSAR Classification Model Based on Weighted Penalized Logistic Regression." *Journal of Chemometrics* 31 doi:10.1002/cem.2915.

Algamal ZY, Lee MH. 2018. "A Two-stage Sparse Logistic Regression for Optimal Gene Selection in High-dimensional microarray data classification." *Advances in Data Analysis and Classification* 13: 753-771 doi:10.1007/s11634-018-0334-1.

Algamal ZY, Lee MH, Al-Fakih AM, Aziz M. 2017. "High-dimensional QSAR Classification Model for Anti-hepatitis C Virus Activity of Thiourea Derivatives Based on the Sparse Logistic Regression Model with a Bridge Penalty." *Journal of Chemometrics* 31: e2889.

Alkhateeb A, Algamal Z. 2020. "Jackknifed Liu-type Estimator in Poisson Regression Model." *Journal of the Iranian Statistical Society* 19:21-37. doi:10.29252/jirss.19.1.21

Asar Y, Arashi M, Wu JJ. 2017. "CiS-S." *Computation Restricted Ridge Estimator in the Logistic Regression Model* 46: 6538-6544.

Asar Y, Genç A. 2015. "New Shrinkage Parameters for the Liu-type Logistic Estimators Communications." *Statistics - Simulation and Computation* 45: 1094-1103 doi:10.1080/03610918.2014.995815.

Batah FSM, Ramanathan TV, Gore SD. 2008. "The Efficiency of Modified Jackknife and Ridge Type Regression Estimators." *A comparison Surveys in Mathematics and its Applications* 3:111 – 122.

Hoerl AE, Kennard RW. 1970. "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics* 12:55-67.

Khurana M, Chaubey YP, Chandra S. 2014. "Jackknifing the Ridge Regression Estimator: A Revisit Communications." *Statistics-Theory and Methods* 43: 5249-5262.

Kibria BG, Månsson K, Shukur GJCE. 2012. *Performance of Some Logistic Ridge Regression Estimators* 40: 401-414.

Kibria BMG. 2003. "Performance of Some New Ridge Regression Estimators Communications." *Statistics - Simulation and Computation* 32: 419-435. doi:10.1081/SAC-120017499

Kibria BMG, Månsson K, Shukur G. 2015. "A Simulation Study of Some Biasing Parameters for the Ridge Type Estimation of Poisson Regression Communications." *Statistics - Simulation and Computation* 44: 943-957. doi:10.1080/03610918.2013.796981

Lukman AF, Emmanuel A, Clement OA, Ayinde K. 2020. "A Modified Ridge-type Logistic Estimator." *Iranian Journal of Science and Technology, Transactions A: Science.* 44:437-443 doi:10.1007/s40995-020-00845-z.

Månsson K, Shukur G. 2011. "A Poisson Ridge Regression Estimator." *Economic Modelling* 28: 1475-1481. doi:10.1016/j.econmod.2011.02.030.

Månsson K, Shukur GJCiS-T, Methods (2011). *On Ridge Parameters in Logistic Regression* 40:3366-3381

Nyquist H. 1988. "Applications of the Jackknife Procedure in Ridge Regression." *Computational Statistics & Data Analysis* 6: 177-183.

Rashad NK, Algamal ZY. 2019. "A New Ridge Estimator for the Poisson Regression Model." *Iranian Journal of Science and Technology, Transactions A: Science* 43: 2921-2928. doi:10.1007/s40995-019-00769-3.

Schaefer R, Roi L, Wolfe RJCiS-T, Methods (1984) A ridge logistic estimator 13:99-113.

Singh B, Chaubey Y, Dwivedi T. 1986. "An Almost Unbiased Ridge Estimator." *Sankhyā: The Indian Journal of Statistics*, Series B 13:342-346

Wu J, Asar YJHJoM, Statistics, 2016. "On Almost Unbiased Ridge Logistic Estimator for the Logistic Regression Model 45:989-998

Yahya Algamal Z. 2018. "Performance of Ridge Estimator in Inverse Gaussian Regression Model." *Communications in Statistics - Theory and Methods* 48: 3836-3849 doi:10.1080/03610926.2018.1481977.

A New Compound Probability Model Applicable to Count Data

Showkat Ahmad Dar*
Department of Statistics, University of Kashmir, Srinagar (J&K), India

Anwar Hassan
Department of Statistics, University of Kashmir, Srinagar (J&K), India

Peer Bilal Ahmad
Department of Mathematical Sciences, Islamic University of Science & Technology, Awantipora, Pulwama (J&K), India

Bilal Ahmad Para
Department of Mathematical Sciences, Islamic University of Science & Technology, Awantipora, Pulwama (J&K), India

In this paper, we obtained a new model for count data by compounding of Poisson distribution with two parameter Pranav distribution. Important mathematical and statistical properties of the distribution have been derived and discussed. Then, parameter estimation is discussed using maximum likelihood method of estimation. Finally, real data set is analyzed to investigate the suitability of the proposed distribution in modeling count data.

Keywords: *Poisson distribution, two parameter Pranav distribution, compound distribution, count data, simulation study, maximum likelihood estimation.*

1. Introduction

There has been a growing concern from the last few decades to obtain flexible parametric probability distributions that can be used to model different types of data sets which cannot be quartered by classical distributions. To obtain such flexible distributions, compounding of probability distribution is comprehensive and advanced technique as it provides a very powerful way to enlarge common

**Corresponding Author: darshowkat2429@gmail.com*

parametric families of distribution to fit data sets that is not adequately fitted by classical probability distributions. Bhati et al. (2015) derived a new generalized Poisson Lindley distribution that finds applications in automobile insurance and epileptic seizure counts. Shaban (1981) built a new compound probability model for analysing count data by compounding Poisson distribution with Inverse Gaussian distribution that finds application in accidents analysis. Hassan S. Bakouch (2018) derived a count data probability model by compounding weighted negative binomial and Lindley distribution. Simon (1955) constructed a new probability model for count data by compounding Poisson with beta distribution. Pielou (1962) obtained a new compound distribution by mixing Poisson with exponential beta distribution. Sankaran (1969) constructed a class of compound Poisson distribution. Rai (1971) presented a compound of Poisson power function distribution. Mahmoudi et al. (2018) introduces a new probability model for count data by compounding Poisson with beta exponential distribution and taking Poisson distribution as parent distribution. Stacy (1962) derived a three parameter life time generalized gamma distribution. Shanker and Fesshaye (2015) introduced a new compounding probability model for count data, by compounding Poisson distribution with Lindley distribution and find its applications in biological science. Aryuyen and Bodhisuwan (2013) obtained a new compound probability model by combining Negative Binomial distribution with generalized exponential distribution. Willmot (1987) introduced the Poisson-inverse Gaussian distribution as an alternative to the negative binomial through compounding machansim. Hassan, Dar and Ahmad (2019) introduced a new compounding probability model for count data, by compounding Poisson distribution with Ishita distribution and find its applications in epileptic seizure. Lord and Geedipall (2011) showed that Poisson distribution tends to under estimate the number of zeros given the mean of the data while the negative Binomial distribution over estimates zero, but under estimate observations with a count. Umeh and Ibenegbu (2019) introduced a two parameter pranav distribution for lifetime data modeling.

In this paper we propose a new count data model which has been built by compounding Poisson distribution with two parameter Pranav distribution and taking Poisson distribution as a parent distribution, as there is a need to find more flexible models for analyzing count data.

2. Definition of Proposed Model (Poisson two parameter Pranav distribution)

If $Z|v \sim P(v)$, where v being itself a random variable following Poisson two parameter Pranav distribution with parameters ζ and η, then determining the distribution that results from marginalizing over v will be known as compound Poison distribution with that of two parameter Pranav distribution, which is denoted by PTPPD (Z; ζ, η). Our proposed model will be discrete as parent distribution is a discrete.

Theorem 1. The probability mass function of a Poisson two parameter Pranav Distribution, i.e., PTPPD (Z; ζ, η) is given by

$$P(Z=z)=\frac{\zeta^4}{(\zeta^4\eta+6)}\left[\frac{\zeta\eta(1+\zeta)^3+(z+3)(z+2)(z+1)}{(1+\zeta)^{z+4}}\right]\ ;\ z=0,1,2,3,...;\zeta,\ \eta>0$$

Proof: The pmf of a Poisson two parameter Pranav distribution can be obtained as

$$j(z\,|\,\nu)=\frac{e^{-\nu}\nu^{x}}{(z)!}\ ;\ z=0,1,2,3,...;\nu>0$$

When its parameter ν follows TPPD with probability density function

$$h(\nu;\zeta)=\frac{\zeta^4(\eta\zeta+\nu^3)e^{-\zeta\nu}}{\eta\zeta^4+6};\nu>0,\zeta,\eta>0$$

The compound of Poisson distribution and two parameter Pranav distribution is given as

$$P(Z=z)=\int_0^\infty g(z\,|\,\nu).h(\nu;\zeta)d\nu$$

$$P(Z=z)=\frac{\zeta^4}{(\zeta^4\eta+6)}\left[\frac{\zeta\eta(1+\zeta)^3+(z+3)(z+2)(z+1)}{(1+\zeta)^{z+4}}\right]$$

$$;z=0,1,2,3,...;\zeta,\eta>0$$

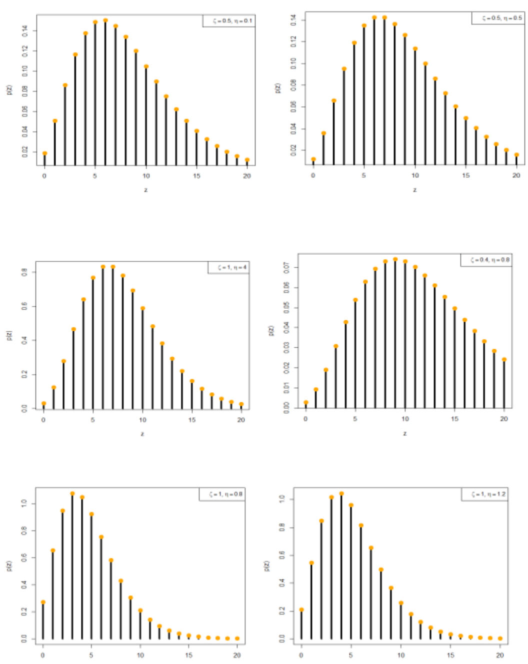

Figure 1 shows the pmf plot for the different values of η and ζ.

The corresponding cdf of Poisson two parameter Pranav distribution is given as

$$F_X(x)=1-\left(\frac{6+24\zeta+6z\zeta+36\zeta^2+21z\zeta^2+3z^2\zeta^2+24\zeta^2+26\zeta^2 z+9z^2\zeta^3+z^3\zeta^3+\eta\zeta^4+3\eta\zeta^5+3\eta\zeta^6+\eta\zeta^7}{(6+\eta\zeta^4)(1+\zeta)^{z+4}}\right)$$

2.1. Random data deneration from Poisson weighted Pranav distribution

In order to simulate the data from PTPPD, we employ the discrete version of inverse cdf method. Simulating a sequence of a random numbers $x_1, x_2, x_3, \ldots, x_n$ from PTPP random variable K with pmf $p(K=x_i)=p_i, \sum_{i=0}^{z} p_{i=1}$ and a cdf $F(K; \zeta, \eta)$, where z may be finite or infinite can be described as following steps:

Step 1: Generate a random number u from uniform distribution $U(0,1)$

Step 2: Generate random number x_i based on

$$if\ u \le p_0 = F(x_0 : \zeta, \eta)\ then\ K = x_0$$

$$f\ p_0 < u \le p_0 + p_1 = F(x_1 : \zeta, \eta)\ then\ K = x_1$$

.

.

.

$$if \sum_{j=0}^{z-1} p_j < u < \sum_{j=0}^{z} p_j = F(x_z : \zeta, \eta)\ then\ K = x_z$$

In order to generate n random numbers $x_1, x_2, x_3, \ldots, x_n$ from PTPPD, repeat step 1 and 2 n times. We have employed R Studio software for running the simulation study of proposed model.

3. Special Case

If we put $\eta = 1$, then Poisson two parameter Pranav distribution reduces to Poisson Pranav Distribution with pmf given as

$$f(z;\eta)=\frac{\zeta^4}{(\zeta^4+6)}\left[\frac{\zeta(1+\zeta)^3+(z+1)(z+2)(z+3)}{(1+\zeta)^{z+4}}\right]$$

4. Reliability Analysis

In this section, we have obtained the reliability and hazard rate function of the proposed PTPPD.

4.1. Reliability Function

$$R(z)=\frac{6+24\zeta+6z\zeta+36\zeta^2+21z\zeta^2+3z^2\zeta^2+24\zeta^2+26\zeta^2z+9z^2\zeta^3+z^3\zeta^3+\eta\zeta^4+3\eta\zeta^5+3\eta\zeta^6+\eta\zeta^7}{(6+\eta\zeta^4)(1+\zeta)^{z+4}}$$

4.2 Hazard Function

$$H.R=\frac{\zeta^4(\zeta\eta(1+\zeta)^3+(z+3)(z+2)(z+1))}{6+24\zeta+6z\zeta+36\zeta^2+21z\zeta^2+3z^2\zeta^2+24\zeta^2+26\zeta^2z+9z^2\zeta^3+z^3\zeta^3+\eta\zeta^4+3\eta\zeta^5+3\eta\zeta^6+\eta\zeta^7}$$

5. Factorial Moment of The Proposed Model

Theorem 5.1. The factorial moments of order s of the proposed model is given by

$$\mu_{(s)'}'=\left[\frac{\zeta^4 s!(\eta\zeta^4)!+(+s+3)(s+2)(s+1)}{(\zeta^4\eta+6)(\zeta^{4+s})}\right]$$

Proof: The sth factorial moment about origin of the PTPPD can be obtained as

$\mu_{(s)}' = E[E(Z^{(s)}| v)$, *where* $Z^{(s)=}Z(Z-1)(Z-2)...(Z-s+1)$

$$\mu_{(s)}'=\int_0^\infty\left[\sum_{z=0}^\infty z^{(s)}\frac{e^{-v}v^z}{(z)!}\right].\frac{\zeta^4(\eta\zeta+v^3)e^{-\zeta v}}{\eta\zeta^4+6}dv$$

$$\mu_{(s)}'=\frac{\zeta^4}{\eta\zeta^4+6}\int_0^\infty\left[v^s\left(\sum_{z=s}^\infty\frac{e^{-v}\lambda^{z-s}}{(z-s)!}\right)\right](\eta\zeta+v^3)e^{-\zeta v}dv$$

Taking $u=z-s$, we get

$$\mu_{(s)}'=\frac{\zeta^4}{\eta\zeta^4+6}\int_0^\infty\left[v^r\left(\sum_{u=0}^\infty\frac{e^{-v}v^u}{u!}\right)\right](\eta\zeta+v^3)e^{-\zeta v}dv$$

$$\mu_{(s)'}'=\left[\frac{\zeta^4 s!(\eta\zeta^4+(+s+3)(s+2)(s+1))}{(\zeta^4\eta+6)(\zeta^{4+s})}\right]$$

6. Recurrence Relation Between Probabilities

If Z~PTPPD (ζ, η) then the pmf of Z is given as

$$P(Z=z)=\frac{\zeta^4}{(\zeta^4\eta+6)}\left[\frac{\zeta\eta(1+\zeta)^3+(z+3)(z+2)(z+1)}{(1+\zeta)^{z+4}}\right]$$

$$P(Z=z+1)=\frac{\zeta^4}{(\zeta^4\eta+6)}\left[\frac{\zeta\eta(1+\zeta)^3+(z+4)(z+3)(z+2)}{(1+\zeta)^{z+5}}\right]$$

$$\frac{P(Z=z+1)}{P(Z=z)}=\frac{\zeta\eta(1+\zeta)^3+(z+4)(z+3)(z+2)}{(1+\zeta)\zeta\eta(1+\zeta)^3+(z+3)(z+2)(z+1)}$$

$$P(Z=z+1)=\frac{\zeta\eta(1+\zeta)^3+(z+4)(z+3)(z+2)}{(1+\zeta)\zeta\eta(1+\zeta)^3+(z+3)(z+2)(z+1)}P(z)$$

7. Estimation of Parameters

In this section, we estimate the unknown parameter of the Poisson two parameter Pranav distribution by using method of maximum likelihood estimation.

7.1. Method of Maximum Likelihood Estimation

Method of Maximum Likelihood Estimation is a simple and the most efficient method of estimation. Let $Z_1, Z_2, Z_3, \ldots, Z_n$, be the random size of sample n drawn from PTPPD, then the likelihood function of PTPPD is given as

$$L(z\mid\zeta,\eta)=\frac{\zeta^{4n}}{(\eta\zeta^4+6)^n}\prod_{i=1}^{n}\left(\frac{(\zeta\eta(1+\zeta)^3+(z+1)(z+2)(z+3))}{(1+\zeta)^{z+4}}\right)$$

$$\log L=4n\log\zeta+\sum_{i=1}^{n}\log(\eta\zeta(1+\zeta)^3+(z+1)(z+2)(z+3))$$
$$-n\log(\eta\zeta^4+6)-(\sum_{i=1}^{n}z_i+4n)\log(1+\zeta)$$

$$\frac{\partial}{\partial\zeta}\log L=\frac{4n}{\zeta}+\sum_{i=1}^{n}\frac{(\eta+6\eta\zeta+12\eta\zeta^2+\eta\zeta^3)}{(\eta\zeta(1+\zeta)^3+(z+1)(z+2)(z+3)}-\frac{3n\eta\zeta^2}{(\eta\zeta^4+6)}-\frac{\sum_{i=1}^{n}z_i+4n}{(1+\zeta)}=0$$

$$\frac{\partial}{\partial\eta}\log L=\sum_{i=1}^{n}\frac{(\zeta(\zeta+1)^3)}{(\eta\zeta(1+\zeta)^3+(z+1)(z+2)(z+3)}-\frac{4\zeta^3}{(\eta\zeta^4+6)}=0$$

The above equations can be solved numerically by using R software 3.5.3 [12].

8. Monte Carlo Simulation

In order to investigate the performance of ML estimators for a finite sample size n using Monte Carlo simulation procedure. Using the inverse cdf method discussed in subsection 2.1, random data is generated from PTPPD. We took four random variable combinations as $\zeta = 2.8$, $\eta = 1.9$, $\zeta = 1.8$, $\eta = 1.2$, $\zeta = 0.5$, $\eta = 0.2$, and $\zeta = 0.2$, $\eta = 0.6$ to carry out the simulation study and the process was repeated 1000 times by going from small to large sample size n = (20, 50, 100, 200, 300 and 500). From Table 1, it is clear that the estimated variance and MSEs when sample size increases. Thus, the agreement between theory and practice improves as the sample size n increases. Hence, the maximum likelihood method performs quite well in estimating the model parameters of Poisson two parameter Pranav distribution.

Table 1. Average Bias, Variance and MSE of ML Estimates of Poisson Two Parameter Pranav Distribution for Different Sample Sizes

n	Parameters	$\zeta = 2.8, \eta = 1.9$				$\zeta = 1.8, \eta = 1.2$			
		Bias	Variance	MSE	Coverage probability	Bias	Variance	MSE	Coverage probability
20	ζ	-0.1212	0.00991	0.024599	0.779	0.065141	0.026776	0.031019	0.911
	η	0.17434	0.091641	0.122035	0.879	0.044127	0.061243	0.080714	0.924
50	ζ	-0.10213	0.006715	0.017145	0.901	-0.00913	0.019104	0.019187	0.929
	η	0.14012	0.061288	0.632513	0.916	0.047141	0.021208	0.021208	0.936
100	ζ	-0.0934	0.005614	0.014337	0.928	0.011207	0.007472	0.007472	0.938
	η	0.07131	0.041271	0.046356	0.931	0.016155	0.000984	0.000984	0.941
200	ζ	-0.0746	0.004124	0.009689	0.941	0.008281	0.000912	0.000912	0.948
	η	-0.0432	0.022131	0.023997	0.949	-0.00925	0.000471	0.000471	0.949
300	ζ	-0.0411	0.001971	0.003660	0.951	0.002914	0.000612	0.000612	0.951
	η	-0.0081	0.000824	0.000824	0.958	0.006714	0.000305	0.000305	0.958
500	ζ	-0.01721	0.000341	0.000341	0.961	0.006923	0.000169	0.000216	0.961
	η	-0.00910	0.000321	0.000321	0.970	0.001247	0.000106	0.000116	0.969
n	Parameters	$\zeta = 0.5, \eta = 0.2$				$\zeta = 0.2, \eta = 0.6$			
		Bias	Variance	MSE	Coverage probability	Bias	Variance	MSE	Coverage probability
20	ζ	0.352110	0.594472	0.718453	0.799	0.439618	1.281363	1.281363	0.891
	η	0.347808	0.393186	0.514156	0.839	0.395411	0.599706	0.756055	0.920
50	ζ	0.141019	0.310896	0.330782	0.906	0.485997	0.458776	0.458776	0.932
	η	0.092191	0.191920	0.200419	0.936	0.368474	0.239788	0.239788	0.939
100	ζ	-0.028951	0.198916	0.199754	0.941	0.246598	0.390818	0.980818	0.943
	η	0.058024	0.146899	0.150265	0.948	0.259943	0.193187	0.193187	0.948
200	ζ	-0.023804	0.108879	0.109446	0.951	0.138508	0.125871	0.145055	0.953
	η	0.003426	0.073616	0.073616	0.954	0.102548	0.094570	0.094570	0.959
300	ζ	0.042858	0.065758	0.065758	0.959	0.038300	0.068572	0.068572	0.964
	η	0.059003	0.039284	0.039284	0.962	0.030150	0.032064	0.032064	0.969
500	ζ	0.342880	0.007762	0.007762	0.968	0.323610	0.003848	0.003848	0.972
	η	0.327908	0.003491	0.003491	0.974	0.295564	0.006709	0.006709	0.979

9. Application of Poisson Two Parameter Pranav Distribution

In order to demonstrate the flexibility and applicability of the proposed distribution in modeling count data set, we have analyzed a data set representing automobile insurance polices (see Klugum et al. 2008), for illustrating the claim that PTPPD is providing better fits when compared to PLD, GD, PD, ZIPD and NBD. The data has a long right tail and approaches to zero slowly. The data sets are given in Table 2.

Table 2. Dataset Representing Automobile Insurance Polices Counts (see Klugman et al. (2008))

Z	0	1	2	3	4	5	6	7	8
Observed Counts	7840	1317	239	42	14	4	4	1	0

For estimation of parameters of the distribution, maximum likelihood method and R software has been used. Parameter estimates, standard errors and model function of the fitted distribution is given in Table 3.

Table 3. Parameter Estimates and Standard Errors for Ffitted Distributions for Dataset 2 (Estimated parameters and standard error for fitted distributions for dataset representing automobile insurance polices counts)

Distribution	Parameter Estimates (Standard Error)	Model function
PTPPD	ζ = 5.62 (0.4) η = 0.08 (0.06)	$P(Z=z)=\frac{\zeta^4}{(\zeta^4\eta+6)}\left[\frac{\zeta\eta(1+\zeta)^3+(z+3)(z+2)(z+1)}{(1+\zeta)^{z+4}}\right]$ $Z=0,1,2,3,\ldots,;\ \zeta\,\eta>0$
PD	ν = 0.21 (0.04)	$p(z)=\frac{e^{-v}v^z}{z!}\quad v>0;\ z=0,1,2,\ldots$
PLD	η = 5.39 (0.11)	$p(z)=\frac{\eta^2(z+\eta+2)}{(\eta+1)^{z+3}}\quad Z=0,1,2,\ldots\theta>0$
GD	p = 0.82 (0.03)	$p(z)=q^z p\quad 0<q<1;\ q=1-p;\ z=0,1,2,\ldots$
NBD	r = 0.70, p = 0.77 (0.2, 0.04)	$p(z)=\binom{z+r-1}{z}p^r q^z,\quad z=0,1,2,\ldots$ $r>0$ and $0<p<1$
ZIPD	η = 0.46, σ = 0.54 (0.02, 0.02)	$p(z)=\begin{cases}\eta+(1-\eta)\frac{e^{-\sigma}\sigma^z}{z!}, & \sigma>0;\ z=0\\ (1-\eta)\frac{e^{-\sigma}\sigma^z}{z!}, & \sigma>0;\ z=0,1,2,..\end{cases}$ $0<\eta<1;\ \sigma>0$

We have fitted Poisson two parameter Pranav distribution (PTPPD), zero inflated Poisson distribution (ZIPD), geometric distribution (GD), Poisson Lindley distribution (PLD), negative binomial distribution (NBD) and Poisson distribution (PD) to the data set given in Table 2. In order to check the goodness of fit of the model and estimation of parameters of the model, Person's chi-square test R studio statistical software has been used. The results are given in Table 4. It is clear from the expected frequencies and the corresponding value of chi-square that Poisson two parameter Pranav distribution provides a satisfactorily better fit for the data set representing automobile insurance claims as compared to other competing models. It is also clear from Figure 2 the values of expected frequencies that Poisson two parameter Pranav distribution provides a closer fit than that provided by other competing models.

Table 4. Fitted PTPPD and Other Competing Models to a Dataset Representing Automobile Insurance Polices

Z	Observed Counts	PD	ZIPD	GD	PLD	NBD	PTPPD
0	7840	627.9	7840	7790.9	7757.7	7879.2	7816.3
1	1317	1703.2	1272.4	1375.25	1381.3	1268.5	1334.6
2	239	2310	296.55	242.75	241.5	248	248.1
3	42	2088.7	46	42.85	41.75	51.3	45.6
4	14	1416.5	5.4	7.55	7.15	10.9	10.1
5	4	768.5	0.5	1.35	1.2	2.4	3.1
6	4	374.4	0.1	0.25	0.2	0.5	2.6
7	1	134.6	0.1	0.1	0.1	0.1	0.5
8	0	64.3	0.1	0.1	0.1	0.1	0.01
Degrees of freedom		4	2	3	3	2	3
Chi-Statistic Value		16517	61.2	23.5	27.4	32.2	3.95
p-value		0	0	0	0	0	0.266

AIC (Akaike information criterion) and BIC (Bayesian information criterion) criterions has been used for comparing our proposed model with other competing models. The lower values of AIC and BIC corresponds to better fitting of model.

As it is clear from Table 5, that the Poisson two parameter Pranav distribution has lesser values of AIC and BIC as compared to other competing models, hence we can concluded that the Poisson two parameter Pranav distribution leads to a better fit than the other competing models for analyzing the data set given in Table 2.

Table 5. AIC, BIC and -logl for Fitted Models to a Dataset Representing Automobile Insurance Polices

Criterion	PD	ZIPD	GD	PLD	NBD	PTPPD
-logl	5359.5	5375.6	5354.7	5356.25	5358	5348.7
AIC	10725	10755.2	10755.2	10714.5	10718	10701.4
BIC	10746.4	10769.5	10769.5	10721.7	10720.2	10701.8

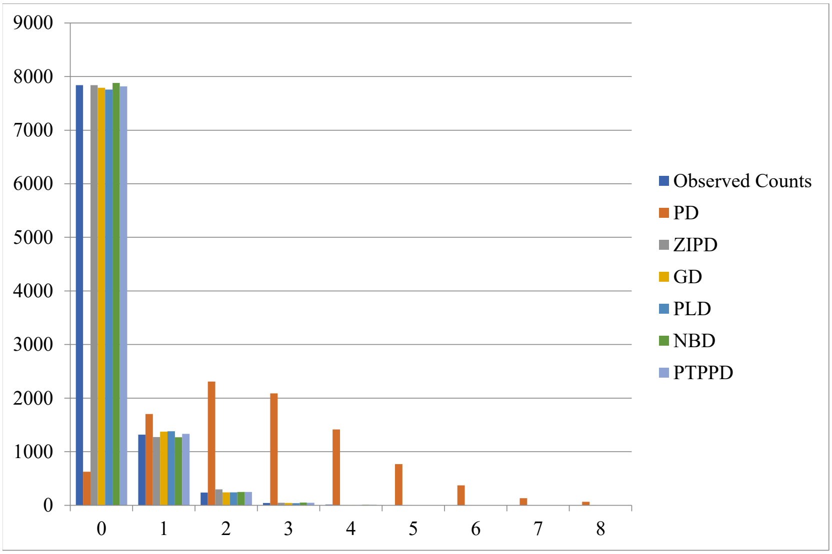

Figure 2. Graphical overview of fitted models to dataset given in Table 2

10. Conclusion

In this paper, we discussed a new model which has been built using compounding technique. Statistical and mathematical properties such as reliability, hazard rate and moments have been discussed. Finally, a real data set is discussed to demonstrate the fitness and applicability of the Poisson two parameter Pranav distribution in modeling count dataset.

References

Aryuyuen, S., & Bodhisuwam, W. 2013. "Negative Binomial Generalized Exponential Distribution." *Applied Mathematical Science* 7(22), 1093-1105.

Bakouch, H. S. (2018). "AWeighted Negative Binomial Lindley Distribution with Applications to Dispersed Data." *Anais da Academia Brasileira de Ciências* 90(3), 2617-2642.

Bhati, D., Sastry, D. V. S., & Qadri, P. M. 2015. "A New Generalized Poisson Lindley Distribution: Applications and Properties." *Austrian Journal of Statistics* 44(4), 35-51.

Hassan, H., Dar, S.A. & Ahmad, P. B. 2019. "Poisson Ishita Distribution: A New Compounding Probability Model." *IOSR Journal of Engineering (IOSRJEN)* 9(2), 38-46.

Lord, D., & Geedipally, S. R. 2011. "The Negative Binomial Lindley Distribution as a Tool for Analyzing Crash Data Characterized by a Large Amount of Zeros." *Accident Analysis & Prevention* 43(5), 1738-1742.

Pielou, E. C. 1962. "Runs of One Species with Respect to Another in Transects Through Plant Populations." *Biometricis* 18(4), 579-593.

Mahmoudi, E., Zamani, H., & Meshkat, R. 2018. "Poisson Beta Exponential Distribution: Properties and Applications." *Journal of Statistical Research of Iran* 15(1), 119-146.

Rai, G. 1971. "A Mathematical Model for Accident Proneness." *Trabajos de Estadistica yde Investigacion Operativa* 22(1-2), 207.

R Core Team 2019. *R Version 3.5.3: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.URL https://www.R-project.org/.

Sankaran, M. 1969. "On Certain Properties of a Class of Compound Poisson Distributions." *Sankhya B* 32, 353-362.

Shaban, S. A. 1981. "On the Discrete Poisson Inverse Gaussian Distribution." *Biometrical Journal* 23(3), 297-303.

Shanker, R., & Hagos, F. 2015. "On Poisson Lindley Distribution and Its Applications to Biological Sciences." *Biometrics & Biostatistics International Journal* 2(4), 1-5.

Simon, P. 1955. "On a Class of Skew Distributions." *Biometrika* 42, 425-440.

Stacy, E. W. 1962. "A Generalization of the Gamma Distribution." *The Annals of Mathematical Statistics* 33(3), 1187-1192.

Umeh, E., and Ibenegbu, A. 2019. "A Two Parameter Pranav Distribution with Properties and its Applications." *Journal of Biostatistics and Epidemiology* 5(1),74-90

Willmot, G. E. 1987. "The Poisson Inverse Gaussian Distribution as an Alternative to the Negative Binomial." *Scandinavian Actuarial Journal* (3-4), 113-127.

Classes of Estimators under New Calibration Schemes using Non-conventional Measures of Dispersion

A. Audu
Department of Mathematics, Usmanu Danfodiyo University Sokoto, Nigeria

R. Singh
Department of Statistics, Banaras Hindu University Varanasi, 221005

S. Khare*
Department of Statistics, Banaras Hindu University Varanasi, 221005

N. S. Dauran
Department of Mathematics, Usmanu Danfodiyo University Sokoto, Nigeria

In this paper, we proposed two classes of estimators under two new calibration schemes for a heterogeneous population by incorporating auxiliary information of Non-Conventional Measures of dispersion which are robust against the presence of outlier in the data.Theoretical results are supported by simulation studies conducted on six bivariate populations generated using exponential and normal distributions. The biases and percentage relative efficiencies (PRE) of the proposed and other related estimators in the study were computed and results indicated that the estimators proposed under suggested calibration schemes perform on average more efficiently than conventional unbiased estimator, Rao and Khan (2016) and Nidhi et al. (2017).

Keywords: *heterogeneous population, Outliers, Estimators, Robust measures, Population mean*

1. Introduction

Traditional method of estimating mean of a study variable y in heterogeneous population stratified into K homogeneous non-overlapping subgroups is to use conventional estimator defined in Eq. (1) as follow:

**Corresponding author: supriya.khare@bhu.ac.in*

$$\tau_{st} = \sum\nolimits_{h=1}^{K} \Psi_h \overline{y}_h \quad (1)$$

where, $\Psi_h = N_h / N$, $\overline{y}_h = n_h^{-1} \sum\nolimits_{i=1}^{n_n} y_{hi}$, n_h is sample size of units drawn with SRSWOR from stratum h, N_h is the size of stratum h and y_{hi} is the i^{th} observation of stratum h.

Utilizing information on supplementary variables to improve the precision of estimators at planning, designing and estimation stage is a well-known approach in sampling theory. Estimation, especially in stratified sampling, entails attaching weight to sample data followed by calculating the weighted mean. Deville and Sarndal (1992) suggested modified weights which improve the precision of an estimate using a procedure called calibration. Many authors have proposed estimators and studied their properties in this direction including Singh & Mohl (1996), Estevao and Sarndal (2000), Audu et al. (2020a), Audu et al. (2020b) and Audu et al. (2021). Tracy et al. (2003) obtained calibration weights for population mean by using first and second order moment of auxiliary variable in stratified random sampling. Kim et al. (2007) utilized calibration approach in defining estimators for population variance in stratified random sampling. Barktus and Pumputis (2010) proposed calibration estimator in stratified sampling for estimating population ratio. Sud et al. (2014) and Estevao & Sarndal (2002) have proposed estimators for different population parameters under different sampling schemes that satisfy the underlying constraints. The weights in stratified sampling are only a function of stratum size which does not gives importance to the stratum information.

Rao and Khan (2016) suggested two new calibration schemes by additively transforming both stratum sample and population means of auxiliary variable using sample and population coefficient of variation respectively in the constraints with respect to usual unbiased estimator $\tau_0 = \sum_{h=1}^{K} \Psi_h \overline{y}_h$, where $\Psi_h = N_h / N$ is the stratum weight and $\overline{y}_h$ is the stratum average of study variable y. The calibration weights Ψ^*_{h1} and Ψ^*_{h2} are selected so as to minimize the distance function $Z_j = \sum_{h=1}^{K} \left(\Psi^*_{hj} - \Psi_h \right)^2 / \Psi_h \phi_h,\ j = 1,2$ subject to calibration constraints $\sum_{h=1}^{K} \Psi^*_{h1} \left(\overline{x}_h + c_{xh} \right) = \sum_{h=1}^{K} \Psi_h \left(\overline{X}_h + C_{Xh} \right)$ and $\sum_{h=1}^{K} \Psi^*_{h2} \left(\overline{x}_h + c_{xh} + 1 \right) = \sum_{h=1}^{K} \Psi_h \left(\overline{X}_h + C_{Xh} + 1 \right)$ respectively, where $\overline{x}_h$ and $\overline{X}_h$ are sample mean and population mean of h^{th} stratum

$$c_{xh}=\frac{s_{xh}}{\bar{x}_h}, C_{Xh}=\frac{S_{Xh}}{\bar{X}_h},, s_{xh}^2=\frac{\sum_{i=1}^{n_h}\left(x_{hi}-\bar{x}_h\right)^2}{n_h-1}, \bar{x}_h=\frac{1}{n_h}\sum_{i=1}^{n_h}x_{hi}, S_{Xh}^2=\frac{\sum_{i=1}^{N_h}\left(x_{hi}-\bar{X}_h\right)^2}{N_h-1}$$

The two schemes proposed are as follow;

$$\tau_{RK1}=\sum_{h=1}^{K}\Psi_h\bar{y}_h\sum_{h=1}^{K}\Psi_h\left(\bar{X}_h+C_X\right)\left(\sum_{h=1}^{K}\Psi_h\left(\bar{x}_h+c_x\right)\right)^{-1} \tag{2}$$

$$\tau_{RK2}=\sum_{h=1}^{K}\Psi_h\bar{y}_h\sum_{h=1}^{K}\Psi_h\left(1+\bar{X}_h+C_X\right)\left(\sum_{h=1}^{K}\Psi_h\left(1+\bar{x}_h+c_x\right)\right)^{-1} \tag{3}$$

where Ψ_h is the stratum weight, C_x is the population coefficient of variation of X, and c_x is the sample coefficient of variation of X.

However, τ_{RK1} and τ_{RK2} are functions of coefficients of variation which can be affected by the presence of extreme values or outliers.

Recently, Nidhi et al. (2017) suggested a new calibration procedure with respect to usual unbiased estimator $\tau_0=\sum_{h=1}^{K}\Psi_h\bar{y}_h$, where $\Psi_h = N_h/N$ is the stratum weight, and $\bar{y}_h$ is the stratum average of study variable y. The calibration weights Ψ_h^* is selected so as to minimize the distance function $Z=\sum_{h=1}^{K}\left(\Psi_h^*-\Psi_h\right)^2/\Psi_h\phi_h$ subject to two calibration constraints $\sum_{h=1}^{K}\Psi_h^*\bar{x}_h=\sum_{h=1}^{K}\Psi_h\bar{X}_h$ and $\sum_{h=1}^{K}\Psi_h^*=1$, where $\bar{x}_h$ and $\bar{X}_h$ are sample mean and population mean of h^{th} stratum. For the cases $\varphi_h=1$ and $\phi_h=\bar{x}_h^{-1}$, Nidhi et al. (2017) obtained new calibrated estimators

$$\tau_{NSSS1}=\sum_{h=1}^{K}\Psi_h\bar{y}_h+\hat{\beta}_{st1}\left(\bar{X}-\sum_{h=1}^{K}\Psi_h\bar{x}_h\right) \tag{4}$$

and

$$\tau_{NSSS2}=\sum_{h=1}^{K}\Psi_h\bar{y}_h+\hat{\beta}_{st2}\left(\bar{X}-\sum_{h=1}^{K}\Psi_h\bar{x}_h\right) \tag{5}$$

respectively where

$$\hat{\beta}_{st1}=\frac{\sum_{h=1}^{K}\Psi_h\bar{x}_h\bar{y}_h-\sum_{h=1}^{K}\Psi_h\bar{y}_h\sum_{h=1}^{K}\Psi_h\bar{x}_h}{\sum_{h=1}^{K}\Psi_h\bar{x}_h^2-\left(\sum_{h=1}^{K}\Psi_h\bar{x}_h\right)^2}$$

and

$$\hat{\beta}_{st2} = \frac{\sum_{h=1}^{K} \Psi_h \bar{y}_h \sum_{h=1}^{K} \Psi_h / \bar{x}_n - \sum_{h=1}^{K} \Psi_h \bar{y}_h / \bar{x}_h}{\sum_{h=1}^{K} \Psi_h \bar{x}_h \sum_{h=1}^{K} \Psi_h / \bar{x}_n - 1}$$

.

2. New Calibration Estimators

The coefficient of variation is affected by outliers, hence, an alternative to the estimators τ_{RK1} and τ_{RK2} would be to replace the coefficient of variation with robust measures of dispersion. Measures of dispersion which are robust to outliers are useful in cases when the population departs from normality. Motivated by Subzar et al. (2018), we proposed new calibration estimators in stratified random sampling using information on robust measures such as Gini's mean difference $G_M(g_M)$, Downton's method $D_M(d_M)$ and probability weighted moments $P_M(p_M)$.

Let $z \in \Re^+$ be population with units z_i, 1,2,…, N, then;

$$G_M(z) = 2N^{-1}(N-1)^{-1} \sum_{i=1}^{N} (2i - N - 1) z_i \quad (6)$$

$$D_M(z) = 2\sqrt{\pi} N^{-1}(N-1)^{-1} \sum_{i=1}^{N} (i - (N+1)/2) z_i \quad (7)$$

$$P_{WM}(z) = \sqrt{\pi} N^{-2} \sum_{i=1}^{N} (2i - (N+1)) z_i \quad (8)$$

Also, let u be sample with unit u_i, 1,2,…, n, , then;

$$g_M(u) = 2n^{-1}(n-1)^{-1} \sum_{i=1}^{n} (2i - n - 1) u_i \quad (9)$$

$$d_M(u) = 2\sqrt{\pi} n^{-1}(n-1)^{-1} \sum_{i=1}^{n} (i - (n+1)/2) u_i \quad (10)$$

$$p_M(u) = \sqrt{\pi} n^{-2} \sum_{i=1}^{n} (2i - (n+1)) u_i \quad (11)$$

Downton's Method, Gini's Mean Method and Probability Weighted Method have been studied by several authors (see David 1968, Downton 1966, Greenwood et al 1979, Yitzhaki 2003). Some existing literature on the improvement of estimators that utilized these robust functions include Abid et al. (2016), Gupta and Yadav (2017) and Yadav and Zaman (2021).

2.1. First new calibration scheme

To obtain the first class of calibration estimator, consider estimator defined in Eq. (12) in stratified sampling;

$$\tau_{ARi} = \sum_{h=1}^{K} \Theta_{hi}^{*} \bar{y}_h, \; i = 1, 2, 3. \quad (12)$$

where Θ^*_{hi} is the new calibration weight that minimizes the Chi-square function denoted Z^* subject to constraints involving the non-standard measures of dispersion, that is,

$$\left.\begin{aligned} &\min \quad Z^* = \sum\nolimits_{h=1}^{K}\left(\Theta^*_{hi} - \Psi_h\right)^2 / \Psi_h\phi_h \\ &s.t \sum\nolimits_{h=1}^{K}\Theta^*_{hi}\left(\bar{x}_h + v_{ih}(x)\right) = \sum_{h=1}^{K}\Psi_h\left(\bar{X}_h + V_{ih}(x)\right) \\ &\quad \sum\nolimits_{h=1}^{K}\Theta^*_{hi} = 1 \end{aligned}\right\} \tag{13}$$

where $\phi_h > 0$ in (13) are suitably chosen weights which determine the form of estimator,

$V_{1h}(x) = G_{Mh}(x), V_{2h}(x) = D_{Mh}(x),\ V_{3h}(x) = P_{Mh}(x),$

$v_{1h}(x) = g_{Mh}(x), v_{2h}(x) = d_{Mh}(x),\ v_{3h}(x) = p_{Mh}(x)$

This minimization problem may be solved by the method of Lagrange multipliers.

Consider the following function

$$L_g = \sum_{h=1}^{K}\frac{\left(\Theta^*_{hi} - \Psi_h\right)^2}{\Psi_h\phi_h} - 2\lambda_1\left(\sum_{h=1}^{K}\Theta^*_{hi}\left(\bar{x}_h + v_{ih}(x)\right) - \sum_{h=1}^{K}\Psi_h\left(\bar{X}_h + V_{ih}(x)\right)\right) - 2\lambda_2\left(\sum\nolimits_{h=1}^{K}\Theta^*_{hi} - 1\right) \tag{14}$$

where λ_j, j =1,2 is Lagrange multiplier. Then, differentiate Lg with respect to $\Theta^*_{hi}, \lambda_1, \lambda_2,$ and equate to 0, that is,

$$\frac{\partial L_g}{\partial \Theta^*_{hi}} = 0,\ \frac{\partial L_g}{\partial \lambda_1} = 0, \frac{\partial L_g}{\partial \lambda_2} = 0 \tag{15}$$

Solving Eq.(15), we get Eq. (16), Eq.(17) and Eq.(18);

$$\Theta^*_{hi} = \Psi_h + \lambda_1\Psi_h\phi_h\left(\bar{x}_h + v_{ih}(x)\right) + \lambda_2\Psi_h\phi_h \tag{16}$$

$$\sum\nolimits_{h=1}^{K}\Theta^*_{hi}\left(\bar{x}_h + v_{ih}(x)\right) - \sum\nolimits_{h=1}^{K}\Psi_h\left(\bar{X}_h + V_{ih}(x)\right) = 0 \tag{17}$$

$$\sum\nolimits_{h=1}^{K}\Theta^*_{hi} - 1 = 0 \tag{18}$$

Substituting the value obtained from Eq. (16) in Eq. (17) and Eq. (18), we get Eq. (19) and Eq. (20) as;

$$\lambda_1 \sum_{h=1}^{K} \Psi_h \phi_h \left(\bar{x}_h + v_{hi}(x)\right)^2 + \lambda_2 \sum_{h=1}^{K} \Psi_h \phi_h \left(\bar{x}_h + v_{hi}(x)\right) = \sum_{h=1}^{K} \Psi_h \left(\bar{X}_h + V_{hi}(x)\right) - \sum_{h=1}^{K} \Psi_h \left(\bar{x}_h + v_{hi}(x)\right) \quad (19)$$

$$\lambda_1 \sum_{h=1}^{K} \Psi_h \phi_h \left(\bar{x}_h + v_{hi}(x)\right) + \lambda_2 \sum_{h=1}^{K} \Psi_h \phi_h = 0 \quad (20)$$

Solving Eq. (19) and Eq. (20) simultaneously, we get expression for λ_1 and λ_2 denoted by λ_1^{opt} and λ_2^{opt} respectively as;

$$\lambda_1^{opt} = \frac{\sum_{h=1}^{K} \Psi_h \phi_h \left(\sum_{h=1}^{K} \Psi_h \left(\bar{X}_h + V_{hi}(x)\right) - \sum_{h=1}^{K} \Psi_h \left(\bar{x}_h + v_{hi}(x)\right)\right)}{\sum_{h=1}^{K} \Psi_h \phi_h \sum_{h=1}^{K} \Psi_h \phi_h \left(\bar{x}_h + v_{hi}(x)\right)^2 - \left(\sum_{h=1}^{K} \Psi_h \phi_h \left(\bar{x}_h + v_{hi}(x)\right)\right)^2} \quad (21)$$

$$\lambda_2^{opt} = -\frac{\sum_{h=1}^{K} \Psi_h \phi_h \left(\bar{x}_h + v_{hi}(x)\right)\left(\sum_{h=1}^{K} \Psi_h \left(\bar{X}_h + V_{hi}(x)\right) - \sum_{h=1}^{K} \Psi_h \left(\bar{x}_h + v_{hi}(x)\right)\right)}{\sum_{h=1}^{K} \Psi_h \phi_h \sum_{h=1}^{K} \Psi_h \phi_h \left(\bar{x}_h + v_{hi}(x)\right)^2 - \left(\sum_{h=1}^{K} \Psi_h \phi_h \left(\bar{x}_h + v_{hi}(x)\right)\right)^2} \quad (22)$$

Now, substituting Eq.(21) and Eq.(22) in Eq.(16), the new calibrated weights Θ_{hi}^* are obtained as

$$\Theta_{hi}^* = \Psi_h + \lambda_1^{opt} \Psi_h \phi_h \left(\bar{x}_h + v_{hi}(x)\right) + \lambda_2^{opt} \Psi_h \phi_h \quad (23)$$

and the new class of calibrated estimators is obtained as;

$$\tau_{ARi} = \sum_{h=1}^{K} \Psi_h \bar{y}_h + \rho_{st}^* \left(\sum_{h=1}^{K} \Psi_h \left(\bar{X}_h + V_{hi}(x)\right) - \sum_{h=1}^{K} \Psi_h \left(\bar{x}_h + v_{hi}(x)\right)\right), \quad i = 1, 2, 3 \quad (24)$$

where

$$\rho_{st}^* = \frac{\sum_{h=1}^{K} \Psi_h \phi_h \sum_{h=1}^{K} \Psi_h \phi_h \left(\bar{x}_h + v_{hi}(x)\right) \bar{y}_h - \sum_{h=1}^{K} \Psi_h \phi_h \bar{y}_h \sum_{h=1}^{K} \Psi_h \phi_h \left(\bar{x}_h + v_{hi}(x)\right)}{\sum_{h=1}^{K} \Psi_h \phi_h \sum_{h=1}^{K} \Psi_h \phi_h \left(\bar{x}_h + v_{hi}(x)\right)^2 - \left(\sum_{h=1}^{K} \Psi_h \phi_h \left(\bar{x}_h + v_{hi}(x)\right)\right)^2}$$

This estimator has estimated mean squared error (MSE) denoted by $M\hat{S}E\left(\tau_{ARi}\right)$ given by;

$$M\hat{S}E\left(\tau_{ARi}\right)=\mathrm{v}\left(\bar{y}_{st}\right)+\rho_{st}^{*^{2}}\,\mathrm{v}\left(\bar{x}_{st}\right)-2\rho_{st}^{*}\,\mathrm{cov}\left(\bar{y}_{st}\bar{x}_{st}\right) \tag{25}$$

where

$$\begin{aligned}
v\left(\bar{y}_{st}\right)&=\sum\nolimits_{h=1}^{K}\Psi_{h}\gamma_{h}S_{yh}^{2},v\left(\bar{x}_{st}\right)\\
&=\sum\nolimits_{h=1}^{K}\Psi_{h}\gamma_{h}S_{xh}^{2},\mathrm{cov}\left(\bar{y}_{st}\bar{x}_{st}\right)\\
&=\sum\nolimits_{h=1}^{K}\Psi_{h}\gamma_{h}\rho_{yxh}S_{yh}S_{xh},\gamma_{h}=\frac{1}{n_{h}}-\frac{1}{N_{h}}
\end{aligned}$$

Further, substituting $\phi_{h}=\left(\bar{x}_{h}+v_{ih}(x)\right)^{-1}$, and $v_{hi}(x)$ be either $g_{Mh}(x)$ or $d_{Mh}(x)$ or $p_{Mh}(x)$ we obtained new estimators as;

$$\left.\begin{aligned}
\tau_{AR1}&=\sum\nolimits_{h=1}^{K}\Psi_{h}\bar{y}_{h}+\rho_{st1}^{*}\left(\sum\nolimits_{h=1}^{K}\Psi_{h}\left(\bar{X}_{h}+G_{Mh}(x)\right)-\sum\nolimits_{h=1}^{K}\Psi_{h}\left(\bar{x}_{h}+g_{Mh}(x)\right)\right)\\
\tau_{AR2}&=\sum\nolimits_{h=1}^{K}\Psi_{h}\bar{y}_{h}+\rho_{st2}^{*}\left(\sum\nolimits_{h=1}^{K}\Psi_{h}\left(\bar{X}_{h}+D_{Mh}(x)\right)-\sum\nolimits_{h=1}^{K}\Psi_{h}\left(\bar{x}_{h}+d_{Mh}(x)\right)\right)\\
\tau_{AR3}&=\sum\nolimits_{h=1}^{K}\Psi_{h}\bar{y}_{h}+\rho_{st3}^{*}\left(\sum\nolimits_{h=1}^{K}\Psi_{h}\left(\bar{X}_{h}+P_{Mh}(x)\right)-\sum\nolimits_{h=1}^{K}\Psi_{h}\left(\bar{x}_{h}+p_{Mh}(x)\right)\right)
\end{aligned}\right\} \tag{26}$$

where

$$\rho_{sti}^{*}=\frac{\sum\nolimits_{h=1}^{K}\Psi_{h}\left(\bar{x}_{h}+v_{hi}(x)\right)^{-1}\sum\nolimits_{h=1}^{K}\Psi_{h}\bar{y}_{h}-\sum\nolimits_{h=1}^{K}\Psi_{h}\left(\bar{x}_{h}+v_{hi}(x)\right)^{-1}\bar{y}_{h}}{\sum\nolimits_{h=1}^{K}\Psi_{h}\left(\bar{x}_{h}+v_{hi}(x)\right)^{-1}\sum\nolimits_{h=1}^{K}\Psi_{h}\left(\bar{x}_{h}+v_{hi}(x)\right)-1},$$

$i=1,2,3$

2.2. Second new calibration scheme

To obtain the second class of the proposed estimators, we let

$$\tau_{ASi}=\sum_{h=1}^{K}\mathrm{H}_{hi}^{*}\bar{y}_{h},\quad i=1,2,3. \tag{27}$$

where H_{hi}^{*} is the new calibration weight such that the Chi-square function T^{*} is defined as

$$\left.\begin{aligned} \min \quad & T^{*}=\sum_{h=1}^{K}\left(\mathrm{H}_{ih}^{*}-\Psi_{h}\right)^{2}/\Psi_{h}\phi_{h} \\ s.t\ & \sum_{h=1}^{K}\mathrm{H}_{ih}^{*}\left(1+\bar{x}_{h}+\nu_{ih}(x)\right)=\sum_{h=1}^{K}\Psi_{h}\left(1+\bar{X}_{h}+V_{ih}(x)\right) \\ & \sum_{h=1}^{K}\mathrm{H}_{ih}^{*}=1 \end{aligned}\right\} \tag{28}$$

Solving for new calibrated weights H_{hi}^{*} using the Lagrange multipliers technique, the new calibrated weights H_{hi}^{*} is

$$\mathrm{H}_{hi}^{*}=\Psi_{h}+\mu_{1}^{opt}\Psi_{h}\phi_{h}\left(1+\bar{x}_{h}+\nu_{hi}(x)\right)+\mu_{2}^{opt}\Psi_{h}\phi_{h}, \tag{29}$$

where

$$\mu_{1}^{opt}=\frac{\sum_{h=1}^{K}\Psi_{h}\phi_{h}\left(\sum_{h=1}^{K}\Psi_{h}\left(1+\bar{X}_{h}+V_{hi}(x)\right)-\sum_{h=1}^{K}\Psi_{h}\left(1+\bar{x}_{h}+\nu_{hi}(x)\right)\right)}{\sum_{h=1}^{K}\Psi_{h}\phi_{h}\sum_{h=1}^{K}\Psi_{h}\phi_{h}\left(1+\bar{x}_{h}+\nu_{hi}(x)\right)^{2}-\left(\sum_{h=1}^{K}\Psi_{h}\phi_{h}\left(1+\bar{x}_{h}+\nu_{hi}(x)\right)\right)^{2}},$$

$$\mu_{2}^{opt}=-\frac{\sum_{h=1}^{K}\Psi_{h}\phi_{h}\left(1+\bar{x}_{h}+\nu_{hi}(x)\right)\left(\sum_{h=1}^{K}\Psi_{h}\left(1+\bar{X}_{h}+V_{hi}(x)\right)-\sum_{h=1}^{K}\Psi_{h}\left(1+\bar{x}_{h}+\nu_{hi}(x)\right)\right)}{\sum_{h=1}^{K}\Psi_{h}\phi_{h}\sum_{h=1}^{K}\Psi_{h}\phi_{h}\left(1+\bar{x}_{h}+\nu_{hi}(x)\right)^{2}-\left(\sum_{h=1}^{K}\Psi_{h}\phi_{h}\left(1+\bar{x}_{h}+\nu_{hi}(x)\right)\right)^{2}}$$

and the new class of calibrated estimators is obtained as:

$$\tau_{ASi}=\sum_{h=1}^{K}\Psi_{h}\bar{y}_{h}+\sigma_{st}^{*}\left(\sum_{h=1}^{K}\Psi_{h}\left(1+\bar{X}_{h}+V_{hi}(x)\right)-\sum_{h=1}^{K}\Psi_{h}\left(1+\bar{x}_{h}+\nu_{hi}(x)\right)\right), \quad i=1,2,3 \tag{30}$$

where

$$\sigma_{st}^{*}=\frac{\sum_{h=1}^{K}\Psi_{h}\phi_{h}\sum_{h=1}^{K}\Psi_{h}\phi_{h}\left(1+\bar{x}_{h}+\nu_{hi}(x)\right)\bar{y}_{h}-\sum_{h=1}^{K}\Psi_{h}\phi_{h}\bar{y}_{h}\sum_{h=1}^{K}\Psi_{h}\phi_{h}\left(1+\bar{x}_{h}+\nu_{hi}(x)\right)}{\sum_{h=1}^{K}\Psi_{h}\phi_{h}\sum_{h=1}^{K}\Psi_{h}\phi_{h}\left(1+\bar{x}_{h}+\nu_{hi}(x)\right)^{2}-\left(\sum_{h=1}^{K}\Psi_{h}\phi_{h}\left(1+\bar{x}_{h}+\nu_{hi}(x)\right)\right)^{2}}$$

The estimated MSE of τ_{ASi} = 1,2,3 denoted by $M\hat{S}E\left(\tau_{ASi}\right)$ is given as:

$$M\hat{S}E\left(\tau_{ASi}\right)=\mathrm{v}\left(\bar{y}_{st}\right)+\sigma_{st}^{*^{2}}\ \mathrm{v}\left(\bar{x}_{st}\right)-2\sigma_{st}^{*}\operatorname{cov}\left(\bar{y}_{st}\bar{x}_{st}\right) \tag{31}$$

Also, substituting $\phi_{h}=\left(1+\bar{x}_{h}+\nu_{ih}(x)\right)^{-1}$, and $\nu_{hi}(x)$ be either $g_{Mh}(x)$ or $d_{Mh}(x)$ or $p_{Mh}(x)$, we obtained new estimators as:

$$\left.\begin{aligned}
\tau_{AS1} &= \sum_{h=1}^{K}\Psi_h\bar{y}_h + \sigma^*_{st1}\left(\sum_{h=1}^{K}\Psi_h\left(1+\bar{X}_h+G_{Mh}(x)\right)-\sum_{h=1}^{K}\Psi_h\left(1+\bar{x}_h+g_{Mh}(x)\right)\right)\\
\tau_{AS2} &= \sum_{h=1}^{K}\Psi_h\bar{y}_h + \sigma^*_{st2}\left(\sum_{h=1}^{K}\Psi_h\left(1+\bar{X}_h+D_{Mh}(x)\right)-\sum_{h=1}^{K}\Psi_h\left(1+\bar{x}_h+d_{Mh}(x)\right)\right)\\
\tau_{AS3} &= \sum_{h=1}^{K}\Psi_h\bar{y}_h + \sigma^*_{st3}\left(\sum_{h=1}^{K}\Psi_h\left(1+\bar{X}_h+P_{Mh}(x)\right)-\sum_{h=1}^{K}\Psi_h\left(1+\bar{x}_h+p_{Mh}(x)\right)\right)
\end{aligned}\right\} \quad (32)$$

where

$$\sigma^*_{sti} = \frac{\sum_{h=1}^{K}\Psi_h\left(1+\bar{x}_h+v_{hi}(x)\right)^{-1}\sum_{h=1}^{K}\Psi_h\bar{y}_h - \sum_{h=1}^{K}\Psi_h\left(1+\bar{x}_h+v_{hi}(x)\right)^{-1}\bar{y}_h}{\sum_{h=1}^{K}\Psi_h\left(1+\bar{x}_h+v_{hi}(x)\right)^{-1}\sum_{h=1}^{K}\Psi_h\left(1+\bar{x}_h+v_{hi}(x)\right)-1},$$

$i = 1,2,3$

2.3. Properties of the new weights Θ^*_{hi} and $\mathrm{H}^*_{hi}, i=1,2,3$

Theorem 1: The proposed weights Θ^*_{hi} and $\mathrm{H}^*_{hi}, i=1,2,3$ are consistent.

Proof: As $n_n \to N_h$, $\bar{x}_h \approx \bar{X}_h$ and $v_{hi}(x) \approx V_{hi}(x)$. Then, the expressions λ_1^{opt} and λ_2^{opt} in $\Theta^*_{hi}, i=1,2,3$ converged to zeros and expressions μ_1^{opt} and μ_2^{opt} in $\mathrm{H}^*_{hi}, i=1,2,3$ tend to zeros. So,

$$\lim_{n_h \to N_h}\frac{\Theta^*_{hi}}{\Psi_h} = 1 \quad (33)$$

$$\lim_{n_h \to N_h}\frac{\mathrm{H}^*_{hi}}{\Psi_h} = 1 \quad (34)$$

Theorem 2: $\lim_{n_h \to N_h}\sum_{h=1}^{K}\Theta^*_{hi} = 1$ and $\lim_{n_h \to N_h}\sum_{h=1}^{K}\mathrm{H}^*_{hi} = 1$.

Proof: Take the summation of Θ^*_{hi} and $\mathrm{H}^*_{hi}, i=1,2,3$ over K, we obtained

$$\sum_{h=1}^{K}\Theta^*_{hi} = 1 + \lambda_1^{opt}\sum_{h=1}^{K}\Psi_h\phi_h\left(\bar{x}_h + v_{hi}(x)\right) + \lambda_2^{opt}\sum_{h=1}^{K}\Psi_h\phi_h \quad (35)$$

$$\sum_{h=1}^{K}\mathrm{H}^*_{hi} = 1 + \mu_1^{opt}\sum_{h=1}^{K}\Psi_h\phi_h\left(1+\bar{x}_h + v_{hi}(x)\right) + \mu_2^{opt}\sum_{h=1}^{K}\Psi_h\phi_h, \quad (36)$$

Take the limits $n_n \to N_h$ of Eqs. (35) and (36), $\lambda_1^{opt} \approx 0, \lambda_2^{opt} \approx 0, \mu_1^{opt} \approx 0, \mu_2^{opt} \approx 0$, $\bar{x}_h \approx \bar{X}_h$, $v_{hi} \approx V_{hi}$, hence the proof.

Theorem 3: $0 < \Theta_{hi}^* < 1$ and $0 < \mathrm{H}_{hi}^* < 1, i = 1,2,3$.

Proof: As $n_n \to N_h$, $\lambda_1^{opt} \approx 0, \lambda_2^{opt} \approx 0, \mu_1^{opt} \approx 0, \mu_2^{opt} \approx 0$, then

$$\lim_{n_h \to N_h} \Theta_{hi}^* = \lim_{n_h \to N_h} \mathrm{H}_{hi}^* = \Psi_h = N_h / N \tag{37}$$

Since $N_h > 0$ (population size of stratum h), $N = \sum_{h=1}^{K} N_h > 0$ (Total population under study) and $N_h <$, N, then $0 < \psi_h < 1$, $\left(\psi_h = \frac{N_h}{N}\right)$, hence the proof.

3. Simulation Study

We conducted simulation studies to examine the performance of the proposed estimators compared to the usual unbiased estimator, Rao and Khan (2016) estimators and Nidhi et al. (2017) estimators. We generated two sets of data of size 1000 units each as the study populations each stratified into three non-overlapping heterogeneous groups of sizes 200, 300 and 500, respectively. The assumptions about the populations are summarized in Table 1. Samples of sizes 20, 30 and 50 respectively from the three strata are obtained 10,000 times by SRSWOR method from each stratum respectively. The biases and precision (PREs) of the considered estimators are computed using Eqs. (38) and (39) respectively.

$$Bias\left(\hat{\theta}\right) = \frac{1}{10000} \sum_{j=1}^{10000} \left(\hat{\theta} - \bar{Y}\right) \tag{38}$$

$$PRE\left(\hat{\theta}_i\right) = \left(\mathrm{var}\left(\theta\right) / \mathrm{var}\left(\theta_l\right)\right)100 \tag{39}$$

where $\mathrm{var}\left(\theta\right) = \frac{1}{10000} \sum_{j=1}^{10000} \left(\tau_{st} - \bar{Y}\right)^2$,

$$\mathrm{var}\left(\hat{\theta}_l\right) = \frac{1}{10000} \sum_{j=1}^{10000} \left(\hat{\theta}_l - \bar{Y}\right)^2, \ \hat{\theta}_l = \tau_{RK1}, \tau_{RK2}, \tau_{AR1}, \tau_{AR2}, \tau_{NSSS1}, \tau_{NSSS2}, \tau_{AR3}, \tau_{AS1}, \tau_{AS2}, \tau_{AS3}$$

Table1. Population used for Empirical Study

Population	**Auxiliary variable *x***	**Study variable *y***
I	$x_h \sim \exp(\lambda_h)$, $\lambda_1=0.2$, $\lambda_2=0.3$, $\lambda_3=0.1$	$y_{hi} = 50\alpha x_{hi} + \xi_{hi}, h = 1,2,3$ $\alpha = 0.5, 1, 1.5, 2.0, 2.5$ $\xi_h \sim N(\phi_h, \psi_h), \phi_h = 0, \psi_h = 1$,
II		$y_{hi} = \alpha x_{hi} + x_{hi}^2 + \xi_{hi}, h = 1,2,3$ $\alpha = 0.5, 1, 1.5, 2.0, 2.5$ $\xi_h \sim N(\phi_h, \psi_h), \phi_h = 0, \psi_h = 1,$
III		$y_{hi} = \alpha x_{hi} + x_{hi}^2 + x_{hi}^3 + \xi_{hi}, h = 1,2,3$ $\alpha = 0.5, 1, 1.5, 2.0, 2.5,$ $\xi_h \sim N(0,1), h = 1,2,3$
IV	$x_h \sim N(\mu_h, \sigma_h)$, $\mu_i = 30$, $\mu_2 = 50$, $\mu_3 = 15$, $\sigma_1 = 25$, $\sigma_2 = 70$, $\sigma_3 = 20$,	$y_{hi} = 50\alpha x_{hi} + \xi_{hi}, h = 1,2,3$ $\alpha = 0.5, 1, 1.5, 2.0, 2.5$ $\xi_h \sim N(\phi_h, \psi_h), \phi_h = 0, \psi_h = 1,$
V		$y_{hi} = \alpha x_{hi} + x_{hi}^2 + \xi_{hi}, h = 1,2,3$ $\alpha = 0.5, 1, 1.5, 2.0, 2.5$ $\xi_h \sim N(\phi_h, \psi_h), \phi_h = 0, \psi_h = 1,$
VI		$y_{hi} = \alpha x_{hi} + x_{hi}^2 + x_{hi}^3 + \xi_{hi}, h = 1,2,3$ $\alpha = 0.5, 1, 1.5, 2.0, 2.5,$ $\xi_h \sim N(0,1), h = 1,2,3$

Table 2. Biases and PREs of the Proposed and Some Existing Related Estimators using Population I

Estimators	Biases					Percentage Relative Efficiencies (PREs)				
	Values of α					Values of α				
	0.5	**1.0**	**1.5**	**2.0**	**2.5**	**0.5**	**1.0**	**1.5**	**2.0**	**2.5**
τ_{st}	-0.1199	-0.2401	-0.3603	-0.4805	0.2993	100.0	100.0	100.0	100.0	100.0
Rao and Khan (2016)										
τ_{RK1}	-0.4781	-0.9558	-1.4336	-1.9113	-1.3249	149.318	149.325	149.329	149.331	150.369
τ_{RK2}	-0.4283	-0.8564	-1.2845	-1.7125	-1.0867	160.017	160.029	160.033	160.036	161.490
Nidhi et al. (2017)										
τ_{NSSS1}	0.0491	0.0982	0.1473	0.1964	1.1046	156.3678	156.3734	156.3748	156.3754	154.9536
τ_{NSSS2}	0.0089	0.0178	0.0266	0.03542	0.9089	158.5629	158.5731	158.5759	158.5772	157.2624
Proposed										
τ_{AR1}	-0.6866	-1.3760	-2.0654	-2.75485	-2.4211	132.6218	132.5576	132.5372	132.5272	133.3487
τ_{AR2}	-0.26467	-0.5279	-0.7912	-1.05451	-0.2559	166.9257	166.9737	166.9902	166.9984	167.4642
τ_{AR3}	-0.3862	-0.7702	-1.1541	-1.5381	-0.8416	163.2251	163.2689	163.2839	163.2915	163.8406
τ_{AS1}	-0.2495	-0.4973	-0.7452	-0.9931	-0.1819	167.2016	167.2560	167.2745	167.2837	167.6974
τ_{AS2}	-0.2520	-0.5024	-0.7528	-1.0032	-0.1955	167.3924	167.4479	167.4667	167.47616	167.947
τ_{AS3}	-0.3797	-0.7570	-1.1342	-1.5113	-0.8140	63.8849	163.9363	163.9538	163.9625	164.5274

Table 3. Biases and PREs of the Proposed and Some Existing Related Estimators using Population II

Estimators	Biases					Percentage Relative Efficiencies (PREs)				
	Values of α					Values of α				
	0.5	**1.0**	**1.5**	**2.0**	**2.5**	**0.5**	**1.0**	**1.5**	**2.0**	**2.5**
τ_{st}	0.0198	0.1082	0.2235	-0.3659	-0.1963	100.0	100.0	100.0	100.0	100.0
Rao and Khan (2016)										
τ_{RK1}	-1.4229	-1.0521	-1.4653	-0.8038	-0.8264	194.0948	225.4986	208.2264	203.2263	209.8469
τ_{RK2}	-1.3801	-1.0133	-1.3950	-0.8333	-0.8324	174.4426	196.9844	184.5767	180.0022	186.5958
Nidhi et al. (2017)										
τ_{NSSS1}	-2.3714	-1.7997	-2.9904	-1.1611	-1.3231	363.4204	475.658	399.9927	368.2536	379.2177
τ_{NSSS2}	-2.2859	-1.7715	-2.8968	-1.1795	-1.2936	355.0888	463.7691	387.9465	360.372	366.1231
Proposed										
τ_{AR1}	1.3128	-2.5101	-1.4748	-0.7958	-1.6216	415.0217	553.2578	493.2792	464.711	498.6249
τ_{AR2}	-1.1827	1.6689	1.6398	0.2355	0.9310	744.0226	915.3678	939.7388	872.3637	865.1686
τ_{AR3}	1.6517	1.9828	1.3296	0.3689	0.5167	748.1287	911.4257	919.9237	861.708	855.4657
τ_{AS1}	-0.7447	1.5009	1.7761	0.6631	1.4878	762.4458	936.2921	989.2003	916.4124	913.2026
τ_{AS2}	-0.8147	1.3728	1.5680	0.6804	1.5195	748.2332	937.1372	965.174	893.2784	889.9554
τ_{AS3}	1.4591	1.1696	1.8547	0.3237	0.6649	747.7217	931.0807	939.6995	877.5081	877.5759

Table 4. Biases and PREs of the Proposed and Some Existing Related Estimators using Population III

Estimators	Biases					Percentage Relative Efficiencies (PREs)				
	Values of α					Values of α				
	0.5	**1.0**	**1.5**	**2.0**	**2.5**	**0.5**	**1.0**	**1.5**	**2.0**	**2.5**
τ_{st}	-85.250	-33.8447	93.9293	-99.575	45.64163	100.0	100.0	100.0	100.0	100.0
Rao and Khan (2016)										
τ_{RK1}	-70.9348	-72.8609	-104.612	-120.712	-47.6903	145.0381	139.5579	135.677	144.9092	187.7669
τ_{RK2}	-66.3908	-70.8230	-95.5546	-112.462	-44.6518	139.6717	135.0246	131.2441	139.5384	185.1077
Nidhi et al. (2017)										
τ_{NSSS1}	-123.729	-99.4131	-192.752	-223.938	-92.5372	194.2402	180.8727	171.6519	191.2706	187.7669
τ_{NSSS2}	-119.598	-97.9878	-184.089	-218.603	-89.8080	191.1841	177.2509	169.5261	188.8559	185.1077
Proposed										
τ_{AR1}	20.0724	84.6589	82.3238	92.8401	41.9473	204.1579	276.4017	277.1548	237.1817	216.0508
τ_{AR2}	-20.8623	-41.1502	-89.8403	-92.0863	-41.7162	210.4178	303.9027	293.7735	273.5902	250.8872
τ_{AR3}	-59.5318	-75.6810	-89.9305	-99.9197	-32.7882	224.7257	302.0734	291.9818	282.2429	263.1409
τ_{AS1}	-23.322	-41.8021	-84.7031	-95.5647	-41.1903	186.9049	301.6926	288.1884	255.8412	226.3107
τ_{AS2}	-22.3187	-39.3112	-41.1733	-43.4198	-27.078	182.7662	295.0898	281.876	249.3989	221.7451
τ_{AS3}	-50.2655	-111.364	-248.983	-205.499	-124.125	196.329	296.5801	284.352	259.5436	234.5219

Table 5. Biases and PREs of the Proposed and Some Existing Related Estimators using Population IV

Estimators	Biases					Percentage Relative Efficiencies (PREs)				
	Values of α					Values of α				
	0.5	**1.0**	**1.5**	**2.0**	**2.5**	**0.5**	**1.0**	**1.5**	**2.0**	**2.5**
τ_{st}	0.2744885	0.5497394	0.8249903	1.100241	1.375492	100	100	100	100	100
Rao and Khan (2016)										
τ_{RK1}	1.7372	3.4735	5.2098	6.9460	8.6823	177.7244	177.7181	177.7161	177.7150	177.7144
τ_{RK2}	1.9543	3.9079	5.8615	7.8151	9.7687	184.4245	184.4309	184.4329	184.4341	184.4347
Nidhi et al. (2017)										
τ_{NSSS1}	2.5830	3.1666	3.7503	4.3339	4.9176	175.632	175.6379	175.6398	175.6408	175.6413
τ_{NSSS2}	2.5477	3.0961	3.6444	4.1928	4.7412	176.6904	176.6964	176.6983	176.6993	176.6999
Proposed										
τ_{AR1}	0.95250	1.9034	2.8544	3.8053	4.7563	179.8097	179.7991	179.7956	179.7939	179.7929
τ_{AR2}	1.8478	3.6959	5.5439	7.392041	9.2401	186.4118	186.4170	186.4187	186.419	186.4200
τ_{AR3}	3.5021	7.00436	10.5066	14.0089	7.5112	178.4900	178.4933	178.4944	178.4949	178.4952
τ_{AS1}	1.8215	3.6434	5.4653	7.2871	9.1089	186.5754	186.5808	186.5822	186.5831	186.5837
τ_{AS2}	1.8432	3.6867	5.5302	7.3737	9.217267	186.6404	186.6455	186.6471	186.6479	186.6485
τ_{AS3}	1.4689	3.9379	5.4069	6.8759	7.3450	78.7401	178.7433	178.7443	178.7448	178.7452

Table 6. Biases and PREs of the Proposed and Some Existing Related Estimators using Population V

Estimators	Biases					Percentage Relative Efficiencies (PREs)				
	Values of α					Values of α				
	0.5	**1.0**	**1.5**	**2.0**	**2.5**	**0.5**	**1.0**	**1.5**	**2.0**	**2.5**
τ_{st}	6.56038	6.550943	6.541506	6.532068	6.522631	100.0	100.0	100.0	100.0	100.0
Rao and Khan (2016)										
τ_{RK1}	13.32939	13.32939	13.32939	13.32939	13.32939	91.30676	91.86576	92.42797	92.9934	93.56203
τ_{RK2}	25.09506	25.09506	25.09506	25.09506	25.09506	95.2063	95.7892	96.3754	96.9649	97.5579
Nidhi et al. (2017)										
τ_{NSSS1}	2.051284	2.002373	1.953462	1.904551	1.85564	214.811	216.178	217.554	218.938	220.329
τ_{NSSS2}	0.8098365	0.754982	0.700128	0.645274	0.590419	215.332	216.691	218.059	219.434	220.817
Proposed										
τ_{AR1}	-36.61819	-36.8667	-37.1151	-37.3636	-37.6121	487.572	486.169	484.733	483.263	481.762
τ_{AR2}	2.630642	2.816348	3.002054	3.187761	3.373467	665.024	664.085	663.020	661.834	660.529
τ_{AR3}	112.4176	113.2183	114.0189	114.8196	115.6202	442.691	441.170	439.603	437.991	436.336
τ_{AS1}	-0.040959	0.138177	0.317312	0.496448	0.67558	795.44	793.314	791.015	788.549	785.924
τ_{AS2}	3.058078	3.248598	3.439118	3.629638	3.820158	655.1993	654.1062	652.8923	651.5614	650.1173
τ_{AS3}	113.2075	114.0165	114.8255	115.6345	116.4435	435.7339	434.1481	432.5177	430.8449	429.1322

Table 7. Biases and PREs of the Proposed and Some Existing Related Estimators using Population VI

Estimators	Biases					Percentage Relative Efficiencies (PREs)				
	Values of α					Values of α				
	0.5	1.0	1.5	2.0	2.5	0.5	1.0	1.5	2.0	2.5
τ_{st}	895.8263	895.8169	895.8075	895.798	895.7886	100.0	100.0	100.0	100.0	100.0
Rao and Khan (2016)										
τ_{RK1}	-5881.96	-5881.96	-5881.96	-5881.96	-5881.96	172.7992	172.8052	172.8113	172.8174	172.824
τ_{RK2}	-4105.46	-4105.46	-4105.46	-4105.46	-4105.46	182.5521	182.5585	182.5649	182.5713	182.578
Nidhi et al. (2017)										
τ_{NSSS1}	-3258.36	-3258.41	-3258.46	-3258.51	-3258.56	175.5434	175.5492	175.5551	175.5609	175.567
τ_{NSSS2}	-3242.38	-3242.44	-3242.49	-3242.55	-3242.61	172.8302	172.8358	172.8414	172.847	172.8525
Proposed										
τ_{AR1}	-7433.06	-7433.31	-7433.55	-7433.80	-7434.05	253.0657	253.0641	253.0626	253.061	253.0595
τ_{AR2}	-6218.15	-6217.97	-6217.78	-6217.59	-6217.41	305.5811	305.6013	305.6216	305.6418	305.6621
τ_{AR3}	3625.091	3625.891	3626.692	3627.493	3628.293	294.7363	294.7562	294.776	294.7959	294.8157
τ_{AS1}	-6197.92	-6197.74	-6197.57	-6197.39	-6197.21	295.649	295.6671	295.6852	295.7033	295.7213
τ_{AS2}	-6210.93	-6210.74	-6210.55	-6210.36	-6210.17	306.1578	306.1782	306.1986	306.219	306.2395
τ_{AS3}	3648.493	3649.302	3650.111	3650.92	3651.729	295.2923	295.3123	295.3324	295.3524	295.3724

4. Discussion

Tables 2, 3, 4, 5, 6 and 7 showed the results of biases and PREs of the usual unbiased, Rao and Khan (2016) and Nidhi et al. (2017) and proposed calibration estimators using populations I, II, III, IV, V and VI respectively defined in Table 1 for different values of α = (0.5, 1.0, 1.5, 2.0, 2.5). The results of the PREs in Table 2 revealed that for all the values of α (coefficients of linear component of response variable models) using linear function, the proposed estimators have highest values except the proposed estimator τ_{AR1} performed below Rao and Khan (2016) and Nidhi et al. (2017) estimators under normal distribution while the results of Table 5 revealed that for all the values of α (coefficients of linear component of response variable models) in the linear function, the proposed estimators have highest values except the proposed estimators τ_{AR1}, τ_{AR2}, τ_{AR3} which performed below Rao and Khan (2016) τ_{RK2} estimator under exponential distribution. Also, the results of the PREs in Tables 3, 4, 6, and 7 revealed that for all the values of α (coefficients of linear component of study (response) variable models) using linear, quadratic and cubic functions in Table 1 for both normal and exponential distributions, the proposed estimators have highest values except some few cases in which the proposed estimators τ_{AS1} and τ_{AS2} performed below Nidhi et al. (2017). These results implied that the proposed estimators on the average are more efficient in estimation of population mean than other related estimators considered in this study.

5. Conclusion

In this study, we used auxiliary character for a heterogeneous population in the form of robust statistical measures based on Gini's mean difference, Downton's method and probability weighted moments. These measures which are not unduly affected by outliers present in the data and provide more efficient estimates of population parameters in the presence of extreme values were used as alternatives for coefficient of variation used by Rao and Khan (2016). From the results of the Tables 2 and 3, it is observed that the estimators proposed under both the calibration schemes are not only robust against outliers but more efficient than usual ratio estimator in stratified sampling.

Acknowledgements

The authors are extremely thankful to refereed for their valuable comments and corrections that helped a lot in the improvement of the paper.

References

Abid, M., Abbas, N., Sherwani, R. A. K., Nazir, H. Z. 2016. "Improved Ratio Estimators for the Population mean using Non-conventional Measures of Dispersion." *Pakistan Journal of Statistics and Operations Research* 12(2): 353-367.

Audu, A., Singh, R. And Khare, S. 2021 "Developing Calibration Estimators for Population Mean Using Robust Measures of Dispersion under Stratified Random." *Statistics in Transition* new series, 22(2): 125–142. DOI: 10.21307/stattrans-2021-019.

Audu, A., Singh, R. V. K., Muhammed, S., Ishaq, O. and Zakari, Y. 2020. "On the Efficiency of Calibration Ratio Estimators of Population Mean**."** *Proceeding of Royal Statistics Society Nigeria Local Group*. 247-261.

Audu, A., Danbaba, A., Abubakar, A., Nakone, B. And Ishaq, O. 2020. "On the Efficiency of Calibration Ratio-Cum-Product Estimators of Population Mean**.** *Proceeding of Royal Statistics Society Nigeria Local Group.* 234-246.

David, H. A. 1968. "Gini's Mean Difference Reconsidered." *Biometrika*, 55, 573-575.

Downton, F. 1966. "Linear Estimates with Polynomial Coefficients." *Biometrika*, 53: 129-141.

Deville, J. C., Sarndal, C. E. 1992. "Calibration Estimators in Survey Sampling." *Journal of American statistical Association*, 87(418): 376-382.

Greenwood, J. A., landwehr, J. M., Matalas, M. C., Wallis, J. S. 1979. "Probability Weighted Moments: Definition and Relation to Parameters of Several Distributions Expressible in Inverse Form." *Water Resour. Res*. 15: 1049-1054.

Gupta, R. K., Yadav, S. K. "New Efficient Estimators of Population Mean using Non-traditional Measures of Dispersion. *Open Journal of Statistics* 7: 394-404.

Singh, A. C., Mohl, C. A. 1996. "Understanding Calibration Estimators in Survey Sampling." *Survey Methodology* 22: 107-111.

Estevao, V.M., Sarndal, C.E. 2000. "A Functional Form Approach to Calibration." *Journal of Official Statistics* 16(4): 379.

Tracy, D.S., Singh, S., Arnab, R. 2003. "Note on Calibration in Stratified and Double Sampling." *Survey Methodology* 29(1): 99-104.

Kim, J.M., Sungur, E.A., Heo, T.Y. 2007. "Calibration Approach Estimators in Stratified Sampling. *Statistics & Probability Letters* 77(1): 99-103.

Barktus, I., Pumputis, D. 2010. "Estimation of Finite Population Ratio Using Calibration of Strata Weights." In *Workshop on Survey Sampling Theory and Methodology*. August 2010: 23-27.

Sud, U.C., Chandra, H., Gupta, V.K. 2014. "Calibration-based Product Estimator in Single- and Two-phase Sampling." *Journal of Statistical Theory and Practice* 8(1): 1-11.

Estevao, V.M., Sarndal, C.E. 2002. "The Ten Cases of Auxiliary Information for Calibration in Two-phase Sampling." *Journal of Official Statistics* 18(2): 233.

Rao, D. K., Tekabu, T., Khan, M. G. 2016. "New Calibration Estimators in Stratified Sampling." *Asia-Pacific World Congress on Computer Science and Engineering*, 66-69.

Nidhi, S., Sisodia, B. V. S., Singh, S., Singh, S. K. 2017. "Calibration Approach Estimation of the Mean in Stratified Sampling and Stratified Double Sampling." *Communications in Statistics-Theory and Methods* 46(10): 4932-4942.

Subzar, M., Maqbool, S., Raja, T. A., Bhat, M. A. 2018. "Estimation of Finite Population Mean in Stratified Random Sampling Using Nonconventional Measures of Dispersion." *Journal of Reliability and Statistical Studies* 11(1): 83-92.

Yitzhaki, S. 2003. "Gini's Mean Difference: A Superior Measure of Variability for Non-normal Distributions." *Metron-International Journal of Statistics* 61: 285-316.

Yadav, S. K., Zaman, T. 2021. "Use of Some Conventional and Nonconventional Parameters for Improving the Efficiency of Ratio-type Estimators." *Journal of Statistics and Management Systems*, 1-21.

Time Series Prediction of CO_2 Emissions in Saudi Arabia Using ARIMA, GM(1,1), and NGBM(1,1) Models

Z. F. Althobaiti
Department of Statistics, Faculty of Science, University of Tabuk,
Universiti Teknologi Malaysia, 81310, Johor Bahru, Johor, Malaysia

A. Shabri
Department of Mathematical Sciences, Faculty of Science,
Universiti Teknologi Malaysia, 81310, Johor Bahru, Johor, Malaysia

The investigation of economic aspects of gas emissions in terms of its volume and consequences is very important, given the current increasing trend. Therefore, the prediction of carbon dioxide emissions in Saudi Arabia becomes necessary. This study uses annual time series data on CO_2 emissions in Saudi Arabia from 1970 to 2016. The study built the prediction model of CO_2 emissions in Saudi Arabia by using Autoregressive Integrated Moving Average (ARIMA), Grey System GM and Nonlinear Grey Bernoulli Model (NGBM), and comparing their efficiency and accuracy based on MAPE metric. The results revealed that Nonlinear Grey Bernoulli Model (NGBM) is more accurate than the other prediction models. The results may be useful to Saudi Arabian government in the development of its future economic policies. As a result, five policy recommendations have been proposed, each of which could play a significant role in the development of acceptable Saudi Arabian climate policies.

Keywords: *annual time series data, Autoregressive Integrated Moving Average (ARIMA), CO_2 emissions, global warming, Grey Model (GM), Nonlinear Grey Bernoulli Model (NGBM), prediction, Saudi Arabia*

1. Introduction

In recent years, one of the major topics on international political plans for global warming has been climate change. This is because of greenhouse gas emissions, mainly CO_2 in the atmosphere (Hossain et al. 2017, Bonga & Chirowa 2014). CO_2 is a type of greenhouse gas (GHGs) emitted due to human activities.

Human activities are among the primary drivers of carbon dioxide emissions, with the most important being the generation of energy from coal, oil, and natural gas, and the use of petroleum products for transportation, aircraft, and vehicle trips.

Saudi Arabia is one of the wealthy oil and industrial nations disposed to carbon dioxide emissions, thus exacerbate global warming. Accordingly, the resulting economic losses from CO_2 emissions are more than those anticipated by the industries. This is in corroboration with the study of Ricke et al. (2018), who estimated that the size of the economic losses that will appear again in the economic results of developing countries, would be greater than their previous benefits from the fossil fuel economy. Nevertheless, the three largest countries that are much concerned of the climate change are the United States, Saudi Arabia, and China, which have been ranked in terms of carbon dioxide emissions.

Another study by Jevrejeva et al. (2018) also warned that failure to reduce greenhouse gas emissions would inevitably lead to sea-level rise, which would have severe economic consequences in the world. For instance, with temperatures reaching pre-industrial levels, floods from sea-level rise could cost society $14 trillion yearly by 2100. Therefore, the prediction of CO_2 emissions, which is the most significant task in time series analysis become necessary. Predictions are extremely essential in many fields such as sciences, economics, agriculture, meteorology, medicine, engineering, and others. The prediction of CO_2 emissions involves predicting the values of the time series from the observed time series. The prediction of CO_2 emissions have become a global concern, as it has shown to assist in raising public knowledge about how to forestall environmental issues (Abdullah & Pauzi 2015). Therefore, to make a realistic estimate of Saudi Arabia's future CO_2 emissions, a fuller understanding of the most suitable prediction models is essential.

Many predictive models, such as ARIMA and gray models have been used by researchers to predict CO_2 emissions. For instance, Nyoni & Bonga, (2019) studied forecasting of CO_2 emissions in India. In the study, ARIMA(2,2,0) model was determined to have the best fit for projecting yearly CO_2 in India for the next 13 years, with an estimate of 3.89 million kt by 2025. Also, Chigora et al. (2019) carried out a research on univariate approach using Box-Jenkins to forecast CO_2 emissions for Zimbabwe's tourism destination vibrancy. The ARIMA(10,1,0) model, which focuses on the amount of carbon dioxide (CO_2) emission in Zimbabwe from 1964 to 2014, was employed to have the most suitable model for forecasting yearly CO_2 emissions for the next 10 years, with the model indicating that it will be around 15,000 kt by 2024. Similarly, Nyoni & Mutongi, (2019) predicted carbon dioxide emissions in China from 1960 to 2017, using autoregressive integrated moving average (ARIMA) models. With a prediction of 10,000,000 kt by 2024, the ARIMA (1, 2, 1) model proved to be the most suitable model for forecasting yearly total CO_2 emissions in China for the next ten years.

Lotfalipour et al. (2013) using the Grey and ARIMA models, estimated that CO_2 emissions in Iran will reach 925.68 million tons in 2020, up to 66% from 2010. Also, employing a differential model to predict CO_2 emissions in Iran, the author used the grey system and Autoregressive Integrated Moving Average, and compared them with the RMSE, MAE, and MAPE metrics models. Based on the findings, the ideal degree of Hannan – Rissanen and Box – Jenkins for ARIMA, the ARIMA(1, 1, 2) model was developed. Even though MAPE metrics for three models were less than 10% accuracy of prediction, the grey system confirmed that the three models demonstrated predicting accuracy. Thus, based on the GM (1, 1) estimates, CO_2 emissions was revealed to reach 925.68 million tons in 2020, representing a 66 percent increase over 2010. Besides, Ho, (2018) has also investigated the grey model.

Chen, (2008) and Chen et al. (2008) termed the recently created Nonlinear Grey Bernoulli Model (NGBM(1, 1)) as precise in handling small time-series datasets with nonlinear variations. Also, in the book published by Liu et al. (2004) termed NGBM(1, 1) as more flexible than the GM(1,1). This is because of the NGBM(1, 1) model's versatility in determining annual unemployment statistics in various nations. This is used to assist governments in developing future labor and economic policy. In 2005, NGBM(1, 1) was also employed to predict the foreign exchange rates of twelve of Taiwan's major trading partners. Both experiments mentioned above revealed that the NGBM(1, 1) could increase the accuracy of the original GM(1,1) simulation and forecasting predictions.

Recently, some researchers attempted to improve the NGBM(1, 1) in various ways, such as Zhou et al. (2009) who used a particle swarm optimization approach to determine the parameter value of "n", and employed the model to predict the power load of the Hubei electric power network. The genetic algorithm was used in (Hsu 2009) to optimize the parameters of the NGBM(1, 1), which was then employed to predict economic developments in Taiwan's integrated circuit industry. Moreover, studies by Xie et al. (2021) projected fuel combustion-related CO_2 emissions using a novel continuous fractional nonlinear grey Bernoulli model with grey wolf optimizer. The study is critical for framing and implementing reasonable plans and policies, owing to diverse national energy structures. Therefore, by simultaneously incorporating conformable fractional accumulation and derivative into the traditional NGBM(1,1) model, it can capture the nonlinear characteristics hidden in sequences. The author thus developed a novel continuous fractional NGBM(1,1) model, dubbed CCFNGBM(1,1), to accurately project CO_2 emissions from fuel combustion in China by 2023. GWO was also used in the study to determine the developing coefficients to enhance the predictability of the newly provided model. However, by replacing the fractional derivative with the integer-order derivative, the model not only improves on the grey forecasting model, but it also provides decision-makers with more dependable forecasts.

The findings of these studies imply that ARIMA, GM (1, 1), and NGBM(1, 1) models has continued to prove to be the most suitable model for predicting yearly CO_2 emissions and could form the underlying basis for predicting CO_2 emissions in Saudi Arabia. In this regard, this study intends to evaluate the accuracy of the predicting models in order to obtain the most precise data prediction.

2. Research Methodology

Three predicting models: ARIMA model, grey model and NGBM(1,1) are used in this study. The reasons why these three models were chosen is firstly due to the ARIMA model, which is a conventional forecasting model that produces more reliable and accurate forecasts. Also, it has the benefit of being able to employ a combination of auto regression, difference, and moving average of different orders to generate the ARIMA(p, d, q) model, which can convey multiple types of information of time series. Secondly, GM(1,1) does not necessitate a large sample size, and the effect of short-term prediction is good. Thirdly, ARIMA model and grey model can be directly compared on the same base. The NGBM(1, 1) is a newly created grey model with wide range of applications in diverse fields. This is due to its precision in handling small time-series datasets with nonlinear variations.

2.1. Autoregressive Integrated Moving Average (ARIMA)

The prediction using ARIMA models statistical method is usually viewed as providing more accurate predictions than econometric methodologies (Song et al. 2003). Also, in terms of forecasting performance, ARIMA models outperformed the multivariate models (Du Preez & Witt 2003). Moreover, ARIMA models outperform naive models and smoothing approaches in terms of overall performance (Goh & Law 2002). ARIMA models were created in the 1970s by Box and Jenkins, and its identification, estimation, and diagnostics method is based on the notion of parsimony (Asteriou & Hall 2015). That is; when the original time series is not stationary, the first order difference process ΔY or second order differences $\Delta^2 Y$, and so on, can be investigated. While, If the differenced process is a stationary process, ARIMA model of that differenced process can be found in practice if differencing is applied, usually $d = 1$, or maybe $d = 2$, is enough. The general form of the ARIMA(p, d, q) can be represented by a backward shift operator as.

$$\phi(B)\,\Delta^d Y_t = \theta(B)\,\varepsilon_t$$

The general autoregressive moving average process with AR order p and MA order q can be written as

$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$ (the p order AR operator)

$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q$ (the p order AR operator)

$\Delta^d = (1 - B)^d$

These processes can be written briefly as: $Y_t \sim$ ARIMA(p, d, q) where ϕ is the autoregressive component's parameter estimate, θ is the moving average component's parameter estimate, Δ is the difference operator, d is the difference, and B is the backshift operator (Box et al. 2015).

2.2. ARIMA model

The ARIMA model is one of the most widely used statistical models for time series forecasts (Box et al. 2015). Its forecast principle is to transfer a non-stationary time series into a stationary time series first. As a result, the dependent variable will be described as a model that only yields its lag value, as well as the actual and lag values of the random error term. The following are the steps in the prediction phase (Wang et al. 2018):

Phase 1: Smooth the time data with a differential tool. In theory, stationarity ensures that the fitted curve formed by sampling time series can continue inertially along the present form in the future, i.e., the data's mean and variance should not be significantly changed.

Phase 2: Create a model that is autoregressive (AR). The autoregressive model is a way of forecasting itself using the variable's historical result data, and it describes the link between current value and previous value. It has the following formula:

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t \tag{1}$$

where y_t represents the current value, μ indicates the constant term, p denotes the order, ϕ_i is the autocorrelation coefficient, and ε_t represents the error.

Phase 3: Create a model based on moving averages (MA). In the autoregressive model, the moving average model concentrates on the accumulation of error components. Random fluctuations in forecasts can be successfully eliminated. It has the following formula:

$$y_t = \mu + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t \tag{2}$$

where θ_i is the MA formula's correlation coefficient.

Phase 4: Create an autoregressive moving average model by combining AR and MA(ARMA). The following is the exact formula. The orders of the autoregressive and moving average models, respectively, are *p* and *q* in this formula. The correlation coefficients of the two models, ϕ_i and θ_i, respectively, must be solved.

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{3}$$

2.3. The Box – Jenkins Methodology

The subjective evaluation of plots of auto-correlograms and partial auto-correlograms of the series is used to identify models in the Box-Jenkins process (Meyler et al. 1998). The initial step in model selection is to vary the series to attain stationarity. The researcher will then assess the correlogram to identify the right sequence of the AR and MA components. Because there are no clear–cut guidelines for determining whether AR and MA components are appropriate. Though, this method of selecting AR and MA components is skewed toward the use of personal judgement. As a result, prior experience is essential. The next step is to estimate the preliminary model, which is followed by diagnostic testing. This is accomplished by creating residuals and analyzing whether they fulfil the parameters of a white noise process, which is common in diagnostic testing. If this is not the case, the model must be re-specified, and the method must be restarted from the second stage. The process may continue indefinitely until a suitable model is produced (Nyoni 2018). This procedure is clearly illustrated in Figure 1.

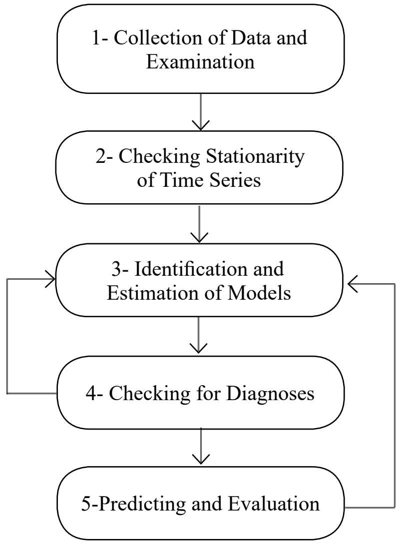

Figure 1. Procedure for ARIMA Forecasting

2.4. Grey Model, GM(1,1)

GM(1,1) denotes a grey forecasting model with one variable and one order. The following is the general steps for creating a grey forecasting model:

Step 1: Create an initial sequence based on observed data.

$$x^{(0)} = (x^{(0)}(1), x^{(0)}(2), \ldots, x^{(0)}(\mathrm{n})) \tag{4}$$

where $x^{(0)}(i)$ denotes the baseline data (state = 0) for the time i

The sample size is n, and the non-negative sequence is $x^{(0)}$. Four data points can be used to develop and build the GM (1, 1) model.

Step 2: Using the initial sequence $x^{(0)}$, to generate the first-order Accumulated Generating Operation (AGO) sequence $x^{(1)}$

$$x^{(1)} = \left(x^{(1)}(1), x^{(1)}(2), \ldots\ldots, x^{(1)}(n)\right) \quad , \quad \mathrm{n} \geq 4 \tag{5}$$

where $x^{(1)}(k)$ is derived as the following formula:

$$x^{(1)}(k) = \sum_{i=1}^{k} x^{(0)}(i) \tag{6}$$

Step 3: Calculate the first-order AGO sequence's mean value:

The average sequences generator's definition is as follows:

$$z^{(1)} = (z^{(1)}(1), z^{(1)}(2), \ldots, z^{(1)}(n))$$

The average value of the sequential data $z^{(1)}(k)$ is define as follows;

$$z^{(1)}(k) = 0.5x^{(1)}(k) + 0.5x^{(1)}(k-1) \quad k = 2, 3, \ldots, n \tag{7}$$

Step 4: Assume the first-order differential equation for the sequence $x^{(1)}$ is as follows:

$$\frac{dx^{(1)}(k)}{dk} + a\, x^{(1)}(k) = b$$

Then its difference equation is shown as:

$$x^{(0)}(k) + a\, z^{(1)}(k) = b \tag{8}$$

where a and b are the estimated parameters of the grey forecasting model.

Step 5: The parameters a and b are calculated using the least-squares method (OLS).

$$\hat{a} = [a,b]^T = \left(B^T B\right)^{-1} B^T Y \tag{9}$$

$$Y = [x^{(0)}(2), x^{(0)}(3),..., x^{(0)}(n)]^T$$

$$B = \begin{bmatrix} -\frac{1}{2}\left(x^{(1)}(1) + x^{(1)}(2)\right) & 1 \\ -\frac{1}{2}\left(x^{(1)}(2) + x^{(1)}(3)\right) & 1 \\ . & \\ . & \\ -\frac{1}{2}\left(x^{(1)}(n-1) + x^{(1)}(n)\right) & 1 \end{bmatrix}$$

Step 6: Under the initial condition $x^{(1)}(1) = x^{(0)}(1)$, the solution of the grey differential equation produces:

$$\hat{x}^{(1)}(k+1) = \left[x^{(0)}(1) - \frac{b}{a}\right]e^{-ak} + \frac{b}{a} \tag{10}$$

Step 7: The first-order inverse accumulated generating operation can be used to get the forecast values $\hat{x}^{(0)}(k+1)$ (IAGO).

$$\hat{x}^{(0)}(k+1) = x^{(1)}(k+1) - x^{(1)}(k) \tag{11}$$

2.5. The Basic NGBM(1,1)

The GM(1,1) method requires obtaining initial data to generate a regular creation sequence for constructing the model. Though, the generative model predicts the original processing data. The nonlinear Bernoulli grey prediction model is based on the GM(1,1) and the differential equation of the modeling to enhance prediction accuracy. This model is commonly utilized by Wang et al. (2011) and Xu et al. (2015). Also, Xie et al. (2021) proposed the Nonlinear Bernoulli Grey Model NBGM(1, 1) to improve prediction accuracy when compared to the original GM (1, 1) model. To achieve this, the following sequence was proposed.

Step 1: Create a starting sequence depending on the data collected.

$$x^{(0)} = (x^{(0)}(1), x^{(0)}(2),\ldots, x^{(0)}(n))$$

where $x^{(0)}(i)$ is the baseline data (state = 0) for time i.

That $x^{(0)}$ is a non-negative sequence, and that n is the sample size. Thus, four data can create and operate a GM (1, 1) model.

Step 2: From the start sequence $x^{(0)}$, generate the first-order Acumulated Generating Operation (AGO) sequence $x^{(1)}$.

$$x^{(1)} = (x^{(1)}(1), x^{(1)}(2), \ldots, x^{(1)}(n)), \quad n \geq 4$$

where $x^{(1)}(k)$ is derived as the following formula:

$$x^{(1)}(k) = \sum_{i=1}^{k} x^{(0)}(i), \quad k = 1, 2, 3, \ldots, n$$

Step 3: Calculate the first-order AGO sequence's mean value.

The following is the definition of the average sequences generator:

$$z^{(1)} = (z^{(1)}(1), z^{(1)}(2), \ldots, z^{(1)}(n))$$

in which $z^{(1)}(k)$ is the background value sequence taken to be the mean generation of consecutive neighbors of $x^{(1)}$ where

$$z^{(1)}(k) = 0.5\, x^{(1)}(k) + 0.5\, x^{(1)}(k-1), \quad k = 2,3,\ldots, n$$

The NGBM(1, 1) model is represented as:

$$x^{(0)}(k) + a\, z^{(1)}(k) = b\, (z^{(1)}(k))^{\gamma}, \gamma \neq 1 \tag{12}$$

which is the whitening equation of the NGBM(1, 1) model.

Step 4: Define the sequence $x^{(1)}$ first-order differential equation is:

$$\frac{dx^{(1)}(k)}{dk} + ax^{(1)}(k) = b(x^{(1)})^{\gamma} \tag{13}$$

The nonlinear parameter γ is given as one, while the linear parameters a and b are determined using the least-squares approach.

Step 5: Assuming the power exponent g is already known, the NGBM(1,1) with the last two parameters are determined as follows:

$$[a, b]^T = (B^T B)^{-1} B^T Y$$

In which T is the matrix transpose. As a result:

$$Y = \left[x^{(0)}(2), x^{(0)}(3), \ldots, x^{(0)}(n) \right]^T$$

$$B = \begin{bmatrix} -z^{(1)}(2) & \left(z^{(1)}(2)\right)^{\gamma} \\ -z^{(1)}(3) & \left(z^{(1)}(3)\right)^{\gamma} \\ \cdot & \\ \cdot & \\ -z^{(1)}(n) & \left(z^{(1)}(n)\right)^{\gamma} \end{bmatrix} \tag{14}$$

Step 6: The following is the solution to the whitening equation:

$$\hat{x}^{(1)}(k+1) = \left\{ \frac{b}{a} + \left[\left(x^{(0)}(1) \right)^{1-\gamma} - \frac{b}{a} \right] e^{-(1-\gamma)ak} \right\}^{\frac{1}{1-\gamma}} \tag{15}$$

Step 7: Compute the original sequence's prediction value:

$$\hat{x}^{(0)}(k) = \hat{x}^{(1)}(k) - \hat{x}^{(1)}(k-1), \quad k = 2, 3, \ldots, m. \tag{16}$$

The NGBM model is a substantial nonlinear grey prediction model in which the power exponent is crucial in grey systems theory. The NGBM model is the GM(1,1) model, especially when $\gamma = 0$. The NGBM model is the grey Verhulst model (GMV) when $\gamma = 2$. Thus, the GM(1,1) and GMV models, in particular, can be considered as versions of the NGBM model. On the other side, the NGBM model can be thought of as a combination of the GM and GMV models. Therefore, the effectiveness of the NGBM model involves specific approaches that may be employed to identify the appropriate power exponent value, which matches the actual data. As a result, the NGBM model can adequately describe the nonlinear properties of real data and improve simulation and prediction accuracy. Wang et al. (2009) used the core principle of information overlap in grey systems to determine the estimated arithmetic of power exponent in the NGBM model. The non-linear programming approach can then be used to calculate the power exponent to minimize mean absolute percentage error (MAPE) (Wang et al. 2012).

2.5.1. Parameter Optimization of the Traditional NGBM(1,1)

The traditional NGBM(1,1) help to determine the expected values for the optimization problem. However, Pao et al. (2012) proposed a relatively simple iterative method for determining the optimal γ.

$$\min_{\gamma} MAPE = \frac{1}{n-1}\sum_{k=2}^{n}\left|\frac{\hat{x}^{(0)}(k)-x^{(0)}(k)}{x^{(0)}(k)}\right| x100\% \quad (17)$$

3. Model Evaluation

The Mean Absolute Percentage Error (MAPE) was used to evaluate the accuracy of the model in this study. This is a widely used criterion for determining the accuracy of predictions. This is presented below:

$$MAPE = \frac{1}{n}\left(\left|\frac{\sum_{i=1}^{n} x_i - \hat{x}_i}{x_i}\right|\right) x100\% \quad (18)$$

where MAPE refers to Mean Absolute Percentage Error, $\hat{x}_i$ is the predicted value, x_i, is the actual value, and the number of data observations n as shown in Table 1.

Table 1. The MAPE Criteria of Prediction Precision

MAPE (%)	≤10	10-20	20-50	≥50
prediction precision	Highly accurate	Good	Reasonable	inaccurate

Source: (Lewis, 1982)

Hence, for a good forecast, the obtained MAPE should be as small as possible (Agrawal & Adhikari, 2013)

4. Results and Discussions

This study is based on 47 yearly CO_2 emissions (kt) observations in Saudi Arabia from 1970 to 2016. The World Bank's online database, which is respected for its trustworthiness and integrity worldwide, provided all the data employed for analysis. The analysis involves using ARIMA, Nonlinear Grey Bernoulli Model (NGBM) and Grey Model (GM) to predict CO_2 emissions. Figure 2 shows that CO_2 emissions (Y) has been increasing from 1970 to 2016, indicating that the trend is not stationary. This implies that the mean and variance of the data are changing over time. Accordingly, the data was divided into two parts: training and testing (forecasting). The data from 2002-2011 was used for training, while the data from 2012 -2016 was used for testing.

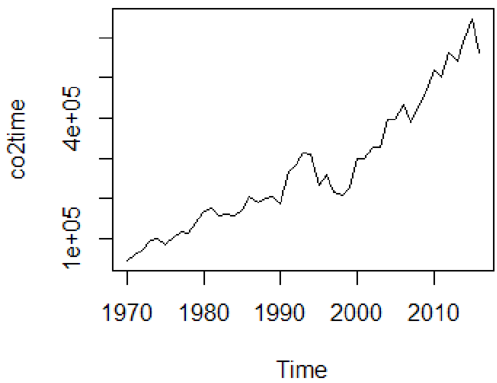

Figure 2. Time series of CO2 emissions in Saudi Arabia

4.1. ARIMA Model

To examine the stationarity of CO_2 emissions, Augmented Dickey- Fuller1 test (1981) was used. According to Table 2, the results of the (ADF) test of the time Series are not stationary in the level at which the calculated statistical significance levels are greater than the level of 0.05. The test results indicated that the time series has reached the stage of stationary after making its first difference. As indicated, the test's statistical significance is less than the 0.05 level.

Table 2. Augmented Dickey- Fuller test (ADF)

Result	**Critical value of ADF**	**The test statistic**	
Non- stationary	-1.7963	0.6552	CO_2
stationary	-3.9973	0.01814	d CO_2

ARIMA(1, 1, 0) with lower AIC is preferable than the one with a higher AIC values (Nyoni 2018). As a result, the ARIMA (1, 1, 0) model is selected as the best as shown in Table 3.

Table 3. Comparison of the Variants of the ARIMA Models

Box-Jenkins Model ARIMA(*p*,*d*,*q*)	AIC
ARIMA(2,1,2)	1097.67
ARIMA(0,1,0)	1094.491
ARIMA(1,1,0)	1093.599
ARIMA(0,1,1)	1094.78
ARIMA(2,1,0)	1095.573
ARIMA(1,1,1)	1095.563
ARIMA(2,1,1)	1098.859

In Table 4, the AR (1) component coefficients are negative and statistically significant at the 5%. This implies that historical CO_2 levels are relevant in describing current and future CO_2 levels in Saudi Arabia. Figure 3 shows that CO_2 emissions in Saudi Arabia are increasing throughout a 13-year period, from 2017 to 2030. Saudi Arabia's CO_2 emissions will reach 747241.6 kt by 2030. As a result, Saudi Arabia will continue to face issues related to global warming and climate change.

Table 4. Results of z Test Coefficients for ARIMA (1,1,0)

variable	coefficient	Standard Error	Z	p-value
AR(1)	-0.2678	0.1547	-1.7312	0.083422
Intercept(mean)	11689.3375	3851.7087	3.0348	0.002407 **

*The *, ** and *** means significant at 10%, 5% and 1% respectively.*

4.2. GM(1,1) and NGBM(1,1) models

The GM(1,1) and NGBM(1,1) models were employed to predict CO_2 emissions in Saudi Arabia. Equation (1) to Equation (6) are used to determine the parameters, develop coefficient *a,* and grey variable *b* for ordinary least squares calculation, and the output is actual GM (1, 1) only variable *a* and *b,* which must be simulated with $\gamma = 0$. The other is determined using the three unknown NGBM(1, 1) variables *a*, *b*, and γ, as given in Table 4. The GRG Nonlinear method of optimization, first devised by Leon Lasdon and Alan Waren, is used to determine the value of the index (Power Exponent γ) (Lasdon et al. 1978). Its implementation as a Fortran software for addressing small to medium-sized issues and some computational findings solved the Nonlinear Optimization Problem. As a result, the value of MAPE was calculated using the NGBM(1,1) at each data point to be predicted by setting the minimum value of MAPE (Pao et al. 2012), and by varying the value of index between -10 and 10 for each data point to be forecasted (Mustaffa & Shabri 2020).

5. Comparative Study

Table 5. Predicted value and MAPE

Year	Actual value	GM(1,1), $\gamma = 0$ $a = -0.0580$, $b = 229.464$		NGBM(1,1), $\gamma = 0.2$ $a = -0.0783$, $b = 315.420$		ARIMA (1,1.0)	
		Predicted VALUE	PE(%)	Predicted VALUE	PE(%)	Predicted VALUE	PE(%)
2002	326.407	314.32	3.70%	299.21	8.33%	305.34	6.45%
2003	327.272	333.11	1.78%	316.38	3.33%	313.46	4.22%
2004	395.834	353.02	10.81%	336.01	15.11%	321.59	18.76%
2005	397.642	374.13	5.91%	358.00	9.97%	329.72	17.08%

2006	432.739	396.49	8.38%	382.35	11.64%	337.84	21.93%
2007	387.777	420.196	8.36%	409.126	5.51%	345.97	10.78%
2008	430.175	445.314	3.52%	438.439	1.92%	354.09	17.69%
2009	468.965	471.934	0.63%	470.433	0.31%	362.22	22.76%
2010	518.491	500.146	3.54%	505.275	2.55%	370.35	28.57%
2011	499.878	530.043	6.03%	543.162	8.66%	378.47	24.29%
MAPE(2000-2011)			4.42%		3.79%		20.82%
2012	564.842	534.679	5.34%	502.516	11.03%	368.6	34.74%
2013	541.047	555.664	2.70%	525.569	2.86%	394.73	27.04%
2014	601.046	577.473	3.92%	552.736	8.04%	402.85	32.98%
2015	647.111	600.137	7.26%	583.585	9.82%	410.98	36.49%
2016	563.449	623.691	10.69%	617.945	9.67%		
MAPE (2012-2016)			5.98%		8.28%		31.37%

Source: Researcher's fieldwork

Table 5 demonstrated that the MAPE value for the NGBM(1,1) in modeling is 3.79%. In comparison, the MAPE value for simulation and forecast data is 8.63%, as shown in Table 5. This implies that the smaller data size influences the MAPE value for simulation data, and its value increases. It is known that the lower the MAPE value, the more accurate the model, and therefore the precise model is at N = 10 for NGBM(1,1).

According to the results, the GM(1,1) has a MAPE of 4.42 %, ARIMA has a MAPE of 20.82%, while NGBM(1,1) has a MAPE of 3.79 %. Compared to the GM(1,1) model and ARIMA model, the NGBM(1,1) model can improve prediction performance. As a result, the prediction value of NGBM(1,1) differs significantly from that of GM(1,1) and ARIMA. This study, therefore, demonstrated that the Mean Absolute Percentage Error (MAPE) is around 3.79% in NGBM(1,1), which implies that the model is about 96.21% the highly accurate in prediction based on the MAPE criteria of prediction precision. While, GM(1,1) is around 4.42% approximately 95.58% highly accurate. But ARIMA(1,1,0) model is around 20.82%, about 79.18% reasonably accurate as presented in the MAPE criteria in Table 1. Consequently, Figure 3 shows the comparison of predictive data of these three models. The NGBM(1,1) model has outperformed than ARIMA(1,1,0) and GM(1,1) model. This is as a result that NGBM(1,1) model has the lower value of MAPE (3.79%) compared with GM(1,1) model (4.42 %) and ARIMA(1,1,0) model (20.82%) . Therefore, NGBM(1,1) delivers the best result among those considered and was used to predict CO_2 emissions in Saudi Arabia. It was also observed that CO_2 emission in Saudi Arabia is continuously increasing as shown in Figure 3. This implies that CO_2 emissions in Saudi Arabia will continue to rise over the next decade from 2017 to 2026, as presented in Figure 4, and the country will face the challenges of global warming, climate change, as well as clean and healthy environment.

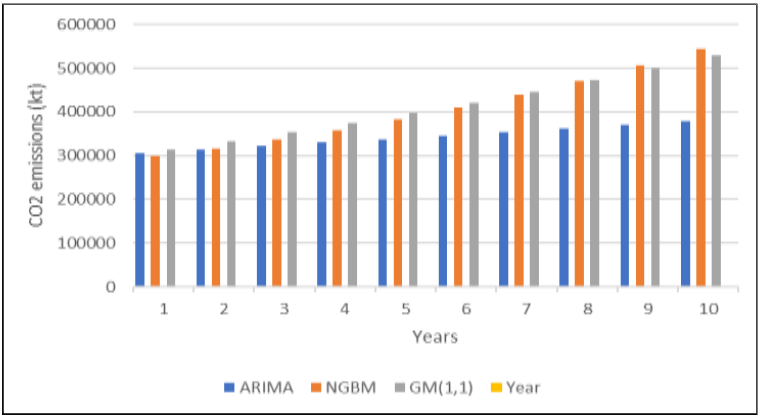

Figure 3. Comparison of predictive data, ARIMA(1,1,0),GAGM(1,1) and GM(1,1) in Saudi Arabia from 2002 to 2011

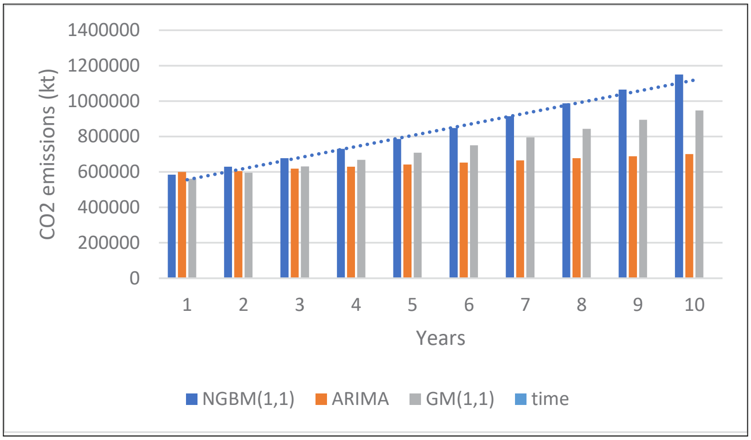

Figure 4. Comparison of predictive data, ARIMA(1,1,0),GAGM(1,1) and GM(1,1) in Saudi Arabia over the next decade from 2017 to 2026

6. Conclusion

This study concluded that NGBM(1,1) modelling is suitable in predicting the future output of the system as it has a high level of accuracy. The prediction accuracy of the NGBM(1,1) model is estimated by Mean Absolute Percentage Error (MAPE). Generally, below 10% MAPE confirms that the NGBM(1,1) provides good prediction accuracy. Therefore, this study shows that NGBM(1,1) is more accurate than ARIMA(1,1,0) and GM(1,1) by evaluating MAPE. The findings of this study are critical for the Saudi government, particularly in terms of medium and long-term economic planning.

To build on these findings and forecast the performance of other sectors, more investigation is recommended. Because this analysis exclusively forecasted CO_2 emissions in Saudi Arabia, this was proposed. CO_2 emissions are influenced by several causes, including the combustion of fossil fuels and the loss of vegetative cover. As a result, humans and ecosystems are affected, and future study will be able to use multi-factor Grey prediction models to develop more precise CO_2 emission projections.

Recommendations

Based on the findings, the following recommendations were made for Saudi Arabia to reach its goal of lowering carbon emissions:

1. Development of renewable energy sources. Although, Saudi Arabia has strong capabilities in solar and winds energy. It does not currently have a competitive sector in the area of renewable energy, so it must be developed.
2. The transition from coal to natural gas.
3. Reliance on nuclear technology to produce energy, which is used in nuclear power plants.
4. There is also a need to keep educating the Saudi people about the need of decreasing pollution levels.
5. The Saudi government should limit pollution by enacting policies such as raising taxes on polluting companies, particularly those that produce fossil fuels, in their everyday operations.

Acknowledgment

The author would like to sincerely thank the University of Tabuk in Saudi Arabia for fully funding this research project and Universiti Teknologi Malaysia (UTM) (University of Studying).

References

Abdullah, L., & Pauzi, H. M. 2015. “Methods in Forecasting Carbon Dioxide Emissions: A Decade Review.” *Jurnal Teknologi* 75(1).

Agrawal, R. K., & Adhikari, R. 2013. *An Introductory Study on Time Series Modeling and Forecasting*. Nova York: CoRR.

Asteriou, D., & Hall, S. G. 2015. *Applied Econometrics*. Macmillan International Higher Education.

Bonga, W. G., & Chirowa, F. 2014. *Level of Cooperativeness of Individuals to Issues of Energy Conservation*. Available at SSRN 2412639.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. 2015. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons.

Chen, C.-I. 2008. "Application of the Novel Nonlinear Grey Bernoulli Model for Forecasting Unemployment Rate. *Chaos, Solitons & Fractals* 37(1): 278–287.

Chen, C.-I., Chen, H. L., & Chen, S.-P. 2008. "Forecasting of Foreign Exchange Rates of Taiwan's Major Trading Partners by Novel Nonlinear Grey Bernoulli Model NGBM(1, 1)." *Communications in Nonlinear Science and Numerical Simulation* 13(6): 1194–1204.

Chigora, F., Thabani, N., & Mutambara, E. 2019. *Forecasting?2 Emission for Zimbabwe's Tourism Destination Vibrancy: A Univariate Approach using Box-Jenkins ARIMA Model*.

Du Preez, J., & Witt, S. F. 2003. "Univariate Versus Multivariate Time Series Forecasting: An Application to International Tourism Demand." *International Journal of Forecasting* 19(3): 435–451.

Goh, C., & Law, R. 2002. "Modeling and Forecasting Tourism Demand for Arrivals with Stochastic Nonstationary Seasonality and Intervention." *Tourism Management* 23(5): 499–510.

Hossain, A., Islam, M. A., Kamruzzaman, M., Khalek, M. A., & Ali, M. A. 2017. *Forecasting Carbon Dioxide Emissions in Bangladesh using Box-Jenkins ARIMA Models*. Department of Statistics, University of Rajshahi.

Hsu, L. C. 2009. "Forecasting the Output of Integrated Circuit Industry Using Genetic Algorithm Based Multivariable Grey Optimization Models." *Expert Systems with Applications* 36(4). https://doi.org/10.1016/j.eswa.2008.11.004

Jevrejeva, S., Jackson, L. P., Grinsted, A., Lincke, D., & Marzeion, B. 2018. "Flood Damage Costs Under the Sea Level Rise With Warming of 1.5 C and 2 C." *Environmental Research Letters* 13(7): 74014.

Lasdon, L. S., Waren, A. D., Jain, A., & Ratner, M. 1978. "Design and Testing of a Generalized Reduced Gradient Code for Nonlinear Programming." *ACM Transactions on Mathematical Software (TOMS)* 4(1): 34–50.

Lewis, C. D. 1982. *Industrial and Business Forecasting Methods: A Practical Guide to Exponential Smoothing and Curve Fitting*. Butterworth-Heinemann.

Liu, S. F., Dang, Y. G., & Fang, Z. G. 2004. *The Theory of Grey System and Its Applications*. Science Press, Beijing.

Lotfalipour, M. R., Falahi, M. A., & Bastam, M. 2013. "Prediction of CO_2 Emissions in Iran Using Grey and ARIMA Models." *International Journal of Energy Economics and Policy* 3(3): 229.

Meyler, A., Kenny, G., & Quinn, T. 1998. *Forecasting Irish inflation using ARIMA models*.

Mustaffa, A. S., & Shabri, A. (2020). An Improved Rolling NGBM(1, 1) Forecasting Model with GRG Nonlinear Method of Optimization for Fossil Carbon Dioxide Emissions in Malaysia and Singapore. *2020 11th IEEE Control and System Graduate Research Colloquium (ICSGRC)*: 32–37.

Nyoni, T. 2018. *Modeling and Forecasting Naira/USD Exchange Rate in Nigeria: A Box-Jenkins ARIMA approach.*Nyoni, Thabani.

Nyoni, T., & Bonga, W. G. 2019. "Prediction of CO_2 Emissions in India Using Arima Models." *DRJ-Journal of Economics & Finance* 4(2): 1–10.

Nyoni, T., & Mutongi, C. 2019. *Modeling and Forecasting Carbon Dioxide Emissions in China Using Autoregressive Integrated Moving Average (ARIMA) Models*.

Pao, H.-T., Fu, H.-C., & Tseng, C.-L. 2012. "Forecasting of CO_2 Emissions, Energy Consumption and Economic Growth in China Using an Improved Grey Model. *Energy* 40(1): 400–409.

Ricke, K., Drouet, L., Caldeira, K., & Tavoni, M. 2018. Country-level Social Cost of Carbon. *Nature Climate Change* 8(10): 895–900.

Song, H., Witt, S. F., & Jensen, T. C. 2003. Tourism Forecasting: Aaccuracy of Alternative Econometric Models. *International Journal of Forecasting* 19(1): 123–141.

Wang, Q., Song, X., & Li, R. 2018. A Novel Hybridization of Nonlinear Grey Model and Linear ARIMA Residual Correction for Forecasting US Shale Oil Production. *Energy* 165: 1320–1331.

Wang, Z.-X., Hipel, K. W., Wang, Q., & He, S.-W. 2011. "An Optimized NGBM(1, 1) Model for Forecasting the Qualified Discharge Rate of Industrial Wastewater in China." *Applied Mathematical Modelling* 35(12): 5524–5532.

Wang, Z. X., Dang, Y. G., Liu, S. F., & Lian, Z. 2009. "Solution of GM (1, 1) Power Model and Its Properties. *Systems Engineering and Electronics* 31(10): 2380–2383.

Wang, Z. X., Dang, Y. G., & Zhao, J. J. 2012. "Optimized GM (1, 1) Power Model and Its Application." *Systems Engineering-Theory & Practice* 32(9): 1973–1978.

Xie, W., Wu, W.-Z., Liu, C., Zhang, T., & Dong, Z. 2021. "Forecasting Fuel Combustion-related CO_2 Emissions by a Novel Continuous Fractional Nonlinear Grey Bernoulli Model with Grey Wolf Optimizer." *Environmental Science and Pollution Research,* 1–17.

Xu, N., Dang, Y., & Cui, J. 2015. "Comprehensive Optimized GM (1, 1) Model and Application for Short term forecasting of Chinese energy consumption and production. *Journal of Systems Engineering and Electronics* 26(4): 794–801.

Zhou, J., Fang, R., Li, Y., Zhang, Y., & Peng, B. 2009. "Parameter Optimization of Nonlinear Grey Bernoulli Model using Particle Swarm Optimization." *Applied Mathematics and Computation* 207(2): 292–299.

Two New Tests for Tail Independence in Extreme Value Models

Mohammad Bolbolian Ghalibaf*
Department of Statistics,
Faculty of Mathematics and Computer Science,
Hakim Sabzevari University, Sabzevar, Iran

This paper proposes two new tests for tail independence in extreme value models. We use the conditional distribution function (df) of $X + Y$, given that $X + Y > c$ based approach of Falk and Michel to test for tail independence in extreme value models. We recommend using Cramer-von Mises and Anderson-Darling tests for tail independence. Simulations show that the two tests are better than the Kolmogorov-Smirnov test which has good results among the proposed tests by Falk and Michel. Finally, by using two real datasets, we illustrate the application of the two proposed tests as well as the traditional tests of Falk and Michel.

Keywords: *extreme value model, tail independence, Copula function, Cramer-von Mises test, Anderson-Darling test, Neyman-Pearson test, Kolmogorov-Smirnov test, Fisher's κ test, Chi-square goodness-of-fit test*

1. Introduction

Tail dependence describes the amount of dependence in the tail of a bivariate distribution. In other words, tail dependence refers to the degree of dependence in the corner of the lower-left quadrant or upper-right quadrant of a bivariate distribution. Definitions of tail dependence for multivariate random vectors are mostly related to their bivariate marginal df's. Geffroy (1958, 1959) and Sibuya (1960) independently introduced the quantity

$$\lambda_u := \lim_{t \to 1^-} P\left(X > F_X^{-1}(t) \mid Y > F_Y^{-1}(t) \right),$$

Corresponding author: m.bolbolian@hsu.ac.ir, m.bolbolian@gmail.com.

where F_X^{-1} and $F_Y^{-1}\mathbb{F}$ are quasi-inverses of F_X and F_Y respectively. This quantity is called the upper tail dependence coefficient provided the limit exists, which is displayed for simplicity as TDC. We say that (X, Y) has upper tail dependence if $\lambda_u > 0$ and upper tail independent or asymptotically independent if $\lambda_u = 0$. Loosely speaking, tail dependence describes the limiting proportion that one margin exceeds a certain threshold given that the other margin has already exceeded that threshold. Several empirical surveys such as An'e and Kharoubi (2003) and Malevergne and Sornette (2004) exhibited that the concept of tail dependence is a useful tool to describe the dependence between extremal data. The TDC can also be defined via the notion of copula. The copula function $C(u,v)$ is a bivariate df with uniform marginals on [0,1], such that $F(x,y) = C(F_X(x), F_Y(y))$. By Sklar's Theorem (Sklar, 1959), this copula exists and is unique if F_X and F_Y are continuous. Also, the copula C is given by $C(u,v) = F(F_X^{-1}(u), F_Y^{-1}(v)),\ \forall u,v \in [0,1]$ (for more details, see Nelsen, 2006). If $C(u,v)$ is the copula of (X, Y), then

$$\lambda_u = \lim_{t \to 1^-} \frac{1-2u+C(u,u)}{1-u}.$$

See Coles et al. (1999). Frahm et al. (2005) introduced estimators for TDC under various assumptions: using a specific distribution, within a class of distributions, using a specific copula function, and within a class of copulas or a nonparametric estimation (without any parametric assumption).

In this paper we restrict our attention to extreme value copulas, i.e., a copula C such that

$$C(u,v) = \exp\left\{ log(uv) A\left(\frac{log(v)}{log(uv)} \right) \right\}, \quad u,v \in [0,1]^2, \tag{1}$$

where, A:[0,1] $\rightarrow$ [1/2,1] is the Pickands dependence function (Pickands 1981). This function is absolutely continuous and convex, satisfies $A(0) = A(1) = 1$, and its derivative has values between –1 and 1. When $A(t) = 1$, Equation (1) yields independence and when in Equation (1) we choose $A(t) = \max\{t, 1-t\}$, then complete dependence obtain. These copulas are useful to model componentwise maxima.

Let (X,Y) be a random vector (rv) with values in $(-\infty,0)^2$, whose df $H(x, y)$ coincides, for $x, y \leq 0$ close to 0, with a max-stable or extreme value df (EV) G with reverse exponential margins, i.e.,

$$G(x, 0) = G(0, x) = exp(x), \quad x \leq 0, \tag{2}$$

and

$$G^n\left(\frac{x}{n},\frac{y}{n}\right)=G(x,y),\quad x,y\le 0, n\in\mathbb{N}.$$

Suppose that $(X_1,Y_1),\ldots,(X_n,Y_n)$ are independent copies of (X,Y). If diagnostic checks of $(X_1,Y_1),\ldots,(X_n,Y_n)$ suggest X,Y to be independent in their upper tail, then modeling with dependencies leads to the over estimation of probabilities of extreme joint events. Some inference problems caused by model mis-specification are, for example, discussed in Dupuis and Tawn (2001). Testing for tail independence is, therefore, mandatory in a data analysis of extreme values.

Falk and Michel (2006) showed that the conditional df of $X+Y$, given that $X+Y>c$, has a limiting df $F(t)=t^2$, $t\in[0,1]$, as c $\uparrow$ 0 if and only if X, Y are tail independent. Otherwise, the limiting df is uniform distribution on [0,1], i.e., $F(t)=t$, $t\in[0,1]$. This result will be utilized to define tests for the tail independence of X, Y which are suggested by the Neyman-Pearson lemma as well as via the goodness-of-fit tests that are based on Fisher's κ, on the Kolmogorov-Smirnov test as well as on the chi-square goodness-of-fit test, applied to the exceedances $X_i+Y_i>c$ in the sample $(X_1,Y_1),\ldots,(X_n,Y_n)$. Using this approach we recommend Cramer-von Mises and Anderson-Darling tests for tail independence.

The organization of the paper is as follows. The next section briefly presents the approach of Falk and Michel (2006) and then expresses their tests for tail independence in extreme value models. Also, we introduce the two proposed tests based on the Cramer-von Mises and Anderson-Darling statistics. Section 3 compares the size and power of the proposed tests as well as the traditional tests for tail independence using Monte Carlo experiments. In Section 4, all tests mentioned in Section 2, are implemented on two real datasets. Finally, conclusions are given in the last section. In this paper, for computation and simulation, we use the R statistical software.

2. Tail Independence Tests

In the following, we assume that the rv (X, Y) has a df $H(x,y)$, which coincides, for $x, y\le 0$ close to 0, with a max-stable or extreme value df (EV) G with reverse exponential margins (Equation (2)). The following theorem from Falk and Michel (2006) is the basis of the tail independence tests in this paper.

Theorem 1. We have uniformly for $t\in[0,1]$ as c $\uparrow$ 0 as

$$P(X+Y>ct\mid X+Y>c)=\begin{cases}t^2(1+O(c)),\ \text{Tail Independence},\\ t(1+O(c)),\ \textit{elsewhere.}\end{cases}$$

Based on this theorem, Falk and Michel (2006) introduced four tests for tail independence in extreme value models, which can be grouped into two different classes: one based on Neyman-Pearson lemma and the other tests based on Fisher's κ, Kolmogorov-Smirnov and chi-square goodness-of-fit tests. These tests are presented below.

2.1. Proposed tests by Falk and Michel

Suppose that $(X_1,Y_1),\ldots,(X_n,Y_n)$ are independent copies of (X,Y). Fix $c < 0$ and consider only those observations X_i,Y_i among the sample that satisfy $X_i + Y_i > c$. Denote these by $C_1, C_2,\ldots, C_{K(n)}$ in the order of their outcome. If c is large enough, then C_i / c, $i = 1, 2,\ldots$ are iid with a common df F_C and are independent of $K(n)$, which is binomial $B(n, q)$ distributed with $q = 1–(1– c)\exp(c)$.

Neyman-Pearson Test. The first test Falk and Michel (2006) introduced is based on Neyman-Pearson lemma. We have to decide, roughly, whether the df of $V_i := C_i / c$, $i = 1, 2,\ldots$ is equal to either the null hypothesis $F_{(0)}(t) = t^2$ or the alternative $F_{(1)}(t) = t$, $0 \leq t \leq 1$. Assuming that these approximations of the df of $V_i := C_i / c$ are exact and that $K(n) = m > 0$, the optimal test for testing $F_{(0)}$ against $F_{(1)}$ is based on the loglikelihood ratio

$$T_{NP} := log\left(\prod_{i=1}^{m}\frac{1}{2V_i}\right) = -\sum_{i=1}^{m} log(V_i) - m\, log(2),$$

if m is large enough, the p-value of this test obtained by using the central limit theorem, that is equal to

$$p_{NP} = \Phi\left(\frac{2\sum_{i=1}^{m} log(V_i) + m}{m^{1/2}}\right),$$

where Φ denotes the df of the standard normal distribution.

The other three tests of Falk and Michel (2006) are goodness-of-fit tests based on C_i / c.

Fisher's κ Test. Conditioning on $K(n) = m > 0$, we consider the rvs

$$U_i := F_c(C_i / c) = \frac{1-(1-C_i)\exp(C_i)}{1-(1-c)\exp(c)}, \quad i = 1,\ldots,m,$$

if X and Y are tail independent and c is close to 0, according to Theorem 1, rvs U_i $(i=1,\ldots,m)$ are iid from uniform distribution on (0,1). Consider the corresponding order statistics $U_{1:m} \leq \ldots \leq U_{m:m}$ and define

$$S_j := U_{j:m} - U_{j\text{-}1:m}, \quad j = 2,\ldots,m,$$

and let $S_1 = U_{1:m}$, $S_{m+1} = 1 - U_{m:m}$. Suppose that

$$M_m := \max_{1 \le j \le m+1} S_j,$$

then, the Fisher's κ test statistic is

$$\kappa_m := (m+1)\, M_m.$$

A table of the critical values of Fisher's κ test is given in Fuller (1976). The p-value of this test is equal to

$$p_\kappa := 1 - G_{m+1}\left(\frac{\kappa_m}{m+1}\right) = 1 - G_{m+1}(M_m),$$

where

$$G_{m+1}(x) = \sum_{j=0}^{m+1} (-1)^j \binom{m+1}{j} (\max(0, 1 - jx))^m, \quad x > 0.$$

Kolmogorov-Smirnov Test. Conditioning on $K(n) = m > 0$, we can apply the Kolmogorov-Smirnov test to rvs U_i ($i = 1, \ldots, m$). Denote $\hat{F}_m(t) := \frac{1}{m}\sum_{i=1}^{m} I_{[0,t]}(U_i)$ be the empirical df of rvs U_i ($i = 1, \ldots, m$), then the Kolmogorov-Smirnov statistic is

$$T_{KS} := m^{1/2} \sup_{t \in [0,1]} |\hat{F}_m(t) - t|.$$

The approximate p-value of Kolmogorov-Smirnov test is equal to

$$p_{KS} := 1 - K(T_{KS}),$$

where K is the df of the Kolmogorov distribution.

Chi-square Test. Conditioning on $K(n) = m > 0$, we can apply the chi-square goodness-of-fit test to rvs U_i ($i = 1, \ldots, m$). For this purpose, we divide the interval [0,1] into k consecutive and disjoint intervals $I_1, \ldots, I_k$ and consider the chi-square statistic

$$\chi^2_{m,k} := \sum_{i=1}^{k} \frac{(m_i - mp_i)^2}{mp_i},$$

where m_i is the number of observations among $U_1, \ldots, U_m$ that fall into the interval I_i and p_i is the length of I_i, $1 \le I \le k$. If m is large, such that for all $i = 1, \ldots, k$ we have $mp_i > 5$, then the statistic $\chi^2_{m,k}$ have chi-square distribution with k–1 degrees of freedom. Therefore, the approximate p-value of this test is equal to

$$p_{\chi^2} := 1 - \chi^2_{k-1}(\chi^2_{m,k}).$$

2.2. The proposed tests

Based on Theorem 1 from Falk and Michel (2006) we propose two new tests for tail independence in extreme value models. These tests are based on Cramer-von Mises and Anderson-Darling statistics.

Cramer-von Mises Test. Conditioning on $K(n) = m > 0$, we can apply the Cramer-von Mises test to rvs U_i $(i = 1, \ldots, m)$. Consider the corresponding order statistics $U_{1:m} \leq \ldots \leq U_{m:m}$, then the Cramer-von Mises statistic is

$$T_{CM} := \frac{1}{12m} + \sum_{i=1}^{m}\left[U_{i:m} - \frac{2i-1}{2m}\right]^2.$$

Csorgo and Faraway (1996) obtained the exact and asymptotic dfs of Cramer-von Mises statistic, where we can use them to calculate p-value of this test. Therefore, approximate p-value of Cramer-von Mises test is equal to

$$p_{CM} := 1 - K(T_{CM}),$$

where K is the df proposed by Csorgo and Faraway (1996).

Anderson-Darling Test. Conditioning on $K(n) = m > 0$, we can apply the Anderson-Darling test to rvs U_i $(i = 1, \ldots, m)$. Consider the corresponding order statistics $U_{1:m} \leq \ldots \leq U_{m:m}$, then the Anderson-Darling statistic is

$$T_{AD} := -m - \frac{1}{m}\sum_{i=1}^{m}(2i-1)[\log(U_{i:m}) + \log(1 - U_{m-i+1:m})].$$

Anderson and Darling (1954) found the limiting df of this statistic. The mean of this limiting df is 1 and the variance is $2(\pi^2 - 9)/3 \sim 0.57974$. Using the limiting df, we can obtain approximate p-value of Anderson-Darling test as below

$$p_{AD} := 1 - A(T_{AD}),$$

where A is the limiting df proposed by Anderson and Darling (1954).

3. Monte Carlo Experiments

In this section, we carried out to evaluate the performance of all above tests for the tail independence by using Monte Carlo experiments. The joint behavior of rv (X,Y) is assumed to be adequately represented by three one-parameter families of extreme value copulas with dependence parameter θ, namely Gumbel copula, Galambos copula and Husler-Reiss copula. Also, we considered Frank copula does not belong to extreme value copulas. The Gumbel copula is defined as

$$C_\theta(u,v) = exp\left\{-\left[(-\ln u)^\theta + (-\ln v)^\theta\right]^{\frac{1}{\theta}}\right\}, \quad \theta \in [1,\infty),$$

Galambos copula is expressed as

$$C_\theta(u,v) = uv\, exp\left\{-\left[(-\ln u)^{-\theta} + (-\ln v)^{-\theta}\right]^{-\frac{1}{\theta}}\right\}, \quad \theta \in [0,\infty),$$

for $\theta \in [0, \infty)$ Husler-Reiss copula is

$$C_\theta(u,v) = exp\left\{\ln u\, \Phi\left(\frac{1}{\theta} + \frac{\theta}{2}\ln\left(\frac{\ln u}{\ln v}\right)\right) + \ln v\, \Phi\left(\frac{1}{\theta} + \frac{\theta}{2}\ln\left(\frac{\ln v}{\ln u}\right)\right)\right\},$$

and Frank copula is specified by

$$C_\theta(u,v) = -\frac{1}{\theta} log\left[1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{(e^{-\theta} - 1)}\right], \quad \theta \in (-\infty,\infty) \setminus \{0\}.$$

For more details about these copulas see Joe (2014).

The Monte Carlo experiments are conducted for the threshold $c = -0.5, -0.1, -0.05$, and based on $K(n) = m = 25$ exceedances under the hypothesis H_0 of the independence of X and Y.

The chi-square statistic uses $k = 4$ intervals of equal length. 10000 replications are performed and we compute the percentage of rejection of H_0. Two characteristics of the tests were of interest: their ability to maintain their nominal level, arbitrarily fixed at 5% throughout the study, and their power under a variety of alternatives. It should be noted that, conditioning on $K(n) = m = 25$, when the threshold c increases to zero, the required sample size increases too.

Tables 1-3 give the percentage of rejection of the hypothesis of the independent tails of X and Y in sampling from different extreme value copulas. In Gumbel, Galambos and Husler-Reiss copulas, the TDC are equal to $2 - 2^{1/\theta}$, $2^{-1/\theta}$ and $2[1-\Phi(1/\theta)]$ respectively. Therefore, in each table, the first row of each test shows the empirical size of the test under the null hypothesis of the tail independence of rv (X,Y) and other rows present the power of these tests under the tail dependence.

Table 1. Percentage of rejection of H_0 by various tests with the underlying Gumbel copula with degrees of dependence θ and 25 exceedances over the threshold c

Test	Dependence Parameter θ	Threshold		
		-0.5	-0.1	-0.05
Neyman-Pearson	1	**0.1550**	**0.0797**	**0.0672**
	2	0.9704	0.9641	0.9703
	5	0.9852	0.9726	0.9698
	10	0.9843	0.9740	0.9701
Fisher's κ	1	**0.0500**	**0.0531**	**0.0494**
	2	0.1991	0.2388	0.2450
	5	0.2290	0.2405	0.2501
	10	0.2299	0.2486	0.2494
Kolmogorov-Smirnov	1	**0.0467**	**0.0515**	**0.0521**
	2	0.6236	0.7267	0.7513
	5	0.7140	0.7485	0.7586
	10	0.7222	0.7542	0.7604
Chi-square	1	**0.0365**	**0.0423**	**0.0407**
	2	0.4720	0.5841	0.6066
	5	0.5682	0.6050	0.6161
	10	0.5750	0.6077	0.6060
Cramer-von Mises	1	**0.0477**	**0.0492**	**0.0536**
	2	0.6841	0.7839	0.8050
	5	0.7702	0.8050	0.8112
	10	0.7742	0.8042	0.8072
Anderson-Darling	1	**0.0468**	**0.0490**	**0.0537**
	2	0.7960	0.8694	0.8879
	5	0.8622	0.8858	0.8893
	10	0.8647	0.8898	0.8913

As seen in tables regardless of the threshold value, except for the Neyman-Pearson test, the size of all tests is close to nominal level 5%, this is shown Bold in Tables 1-3. Of course, by choosing the small threshold close to 0 we ensure that the size of the Neyman-Pearson test also controls. This is inspected in Lemma 3.1 of Falk and Michel (2006).

Table 2. Percentage of rejection of H_0 by various tests with the underlying Galambos copula with degrees of dependence θ and 25 exceedances over the threshold c

Test	Dependence Parameter θ	Threshold		
		-0.5	-0.1	-0.05
Neyman-Pearson	0	**0.1688**	**0.0906**	**0.0917**
	2	0.9805	0.9674	0.9713
	5	0.9856	0.9713	0.9708
	10	0.9853	0.9729	0.9721
Fisher's κ	0	**0.0485**	**0.0528**	**0.0523**
	2	0.2104	0.2351	0.2424
	5	0.2304	0.2415	0.2460
	10	0.2335	0.2386	0.2415
Kolmogorov-Smirnov	0	**0.0510**	**0.0500**	**0.0498**
	2	0.6742	0.7392	0.7571
	5	0.7132	0.7453	0.7535
	10	0.7165	0.7509	0.7557
Chi-square	0	**0.0434**	**0.0400**	**0.0368**
	2	0.5266	0.5938	0.6083
	5	0.5758	0.6064	0.6130
	10	0.5671	0.6100	0.6119
Cramer-von Mises	0	**0.0536**	**0.0502**	**0.0523**
	2	0.7282	0.7918	0.8058
	5	0.7698	0.8013	0.8068
	10	0.7698	0.8106	0.8063
Anderson-Darling	0	**0.0550**	**0.0527**	**0.0545**
	2	0.8306	0.8771	0.8896
	5	0.8616	0.8878	0.8886
	10	0.8622	0.8873	0.8872

Comparison of the power of the tests shows that the Neyman-Pearson test having the largest power followed by the Anderson-Darling, Cramer-von Mises, Kolmogorov-Smirnov and chi-square tests, respectively.

Table 3. Percentage of rejection of H_0 by various tests with the underlying Husler-Reiss copula with degrees of dependence θ and 25 exceedances over the threshold *c*

Test	Dependence Parameter θ	Threshold		
		-0.5	-0.1	-0.05
Neyman-Pearson	0	**0.1652**	**0.0737**	**0.0633**
	2	0.9774	0.9716	0.9705
	5	0.9847	0.9700	0.9701
	10	0.9870	0.9723	0.9684
Fisher's κ	0	**0.0487**	**0.0507**	**0.0497**
	2	0.1974	0.2348	0.2509
	5	0.2251	0.2485	0.2496
	10	0.2288	0.2438	0.2421
Kolmogorov-Smirnov	0	**0.0484**	**0.0497**	**0.0522**
	2	0.6602	0.7382	0.7509
	5	0.7118	0.7464	0.7556
	10	0.7245	0.7398	0.7577
Chi-square	0	**0.0373**	**0.0389**	**0.0391**
	2	0.5111	0.5895	0.6047
	5	0.5603	0.6049	0.6119
	10	0.5810	0.5994	0.6121
Cramer-von Mises	0	**0.0526**	**0.0485**	**0.0532**
	2	0.7186	0.7886	0.8013
	5	0.7641	0.8000	0.8067
	10	0.7801	0.7984	0.8155
Anderson-Darling	0	**0.0512**	**0.0496**	**0.0524**
	2	0.8234	0.8774	0.8846
	5	0.8599	0.8832	0.8850
	10	0.8684	0.8811	0.8885

As Falk and Michel (2006) pointed out the distribution of p_κ is almost not affected, therefore the test for the independence of *X* and *Y* based on Fisher's κ fails. These results are viewable in Tables 1-3.

Table 4. Percentage of rejection of H_0 by various tests with the underlying Frank copula with degrees of dependence θ and 25 exceedances over the threshold c

Test	Dependence Parameter θ	Threshold		
		-0.5	-0.1	-0.05
Neyman-Pearson	0	0.1589	0.0739	**0.0621**
	2	0.3928	0.1094	**0.0804**
	5	0.6683	0.1626	**0.1053**
	10	0.8722	0.2726	**0.1564**
Fisher's κ	0	0.0502	0.0479	**0.0471**
	2	0.0655	0.0547	**0.0512**
	5	0.1021	0.0548	**0.0552**
	10	0.1550	0.0703	**0.0525**
Kolmogorov-Smirnov	0	0.0502	0.0494	**0.0447**
	2	0.0997	0.0572	**0.0491**
	5	0.2434	0.0737	**0.0557**
	10	0.4726	0.1161	**0.0729**
Chi-square	0	0.0403	0.0424	**0.0374**
	2	0.0664	0.0437	**0.0372**
	5	0.1592	0.0519	**0.0433**
	10	0.3329	0.0760	**0.0518**
Cramer-von Mises	0	0.0521	0.0491	**0.0439**
	2	0.1086	0.0577	**0.0507**
	5	0.2829	0.0763	**0.0584**
	10	0.5252	0.1322	**0.0786**
Anderson-Darling	0	0.0505	0.0507	**0.0458**
	2	0.1161	0.0604	**0.0499**
	5	0.3050	0.0783	**0.0577**
	10	0.5660	0.1389	**0.0806**

Table 4 illustrates the percentage of rejection of the hypothesis of the independent tails of X and Y in sampling from Frank copula. In Frank copula, for all values of the dependence parameter θ, TDC is equal to zero; i.e. X and Y are tail independent. Therefore, this table shows the empirical size of the test under the null hypothesis of the tail independence of rv (X,Y). As seen in Table 4, when the dependence parameter θ is zero (i.e. data does not have any dependency), except for the Neyman-Pearson test, the size of all tests is close to nominal level 5% and by choosing the small threshold the size of the Neyman-Pearson test also controls. By increasing the dependence parameter, although X and Y do not have tail dependence, the empirical size of the tests are violated. Looking at Table

4, we observe that in this case if the threshold value is close to 0, the empirical level approaches the nominal level, this is shown Bold in Table 4. The results of Table 4 show that, even if rv (*X*,*Y*) does not belong to extreme value model, tail independence tests for a small threshold still have good performance.

4. Data Analysis

In this section, the application of tail independence tests is illustrated using two different datasets. The first one is due to Cornwell and Trumbull (1994), who prepared based on the transcript of crime in North Carolina regarding 24 variables. The dataset included a panel of 90 observational units (counties) from 1981 to 1987, i.e. total number of observations is 630. We consider the two variables density (people per square mile) and crmrte (crimes committed per person) and other variables are ignored. We consider this dataset as Crime data. The second dataset, reported from "Investing.com." This site is a global financial portal and internet brand composed of 28 editions in 21 languages and mobile apps for Android and iOS that provide news, analysis, streaming quotes and charts, technical data and financial tools about the global financial markets. We consider stock price pairs from two Japanese multinational automaker: Honda Motor and Mazda Motor. Our sample period covers a total 758 observations from 10 Sep. 2014 to 16 Oct. 2017. We call this dataset as Stock data. In Figure 1, we draw scatter plots of empirical df of pairs for two datasets.

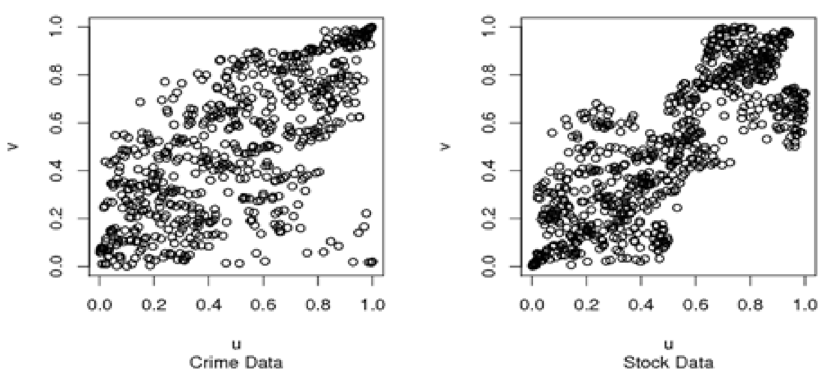

Figure 1. Scatter Plots of Empirical df of Pairs

We use a specific copula method for estimating TDC. For this purpose, we fitted three famous Archimedean copulas to the two datasets and obtained Cramer-von Mises statistic $S_n^{(B)}$ introduced by Genest et al. (2009), where is based on Rosenblatt's transform. It should be noted that the margins are estimated by empirical dfs. The results are shown in Table 5.

Table 5. Copula goodness-of-fit test for two datasets

Copula under H_0	Crime Data				Stock Data			
	p.value	AIC	$\hat{\theta}$	TDC	p.value	AIC	$\hat{\theta}$	TDC
Clayton	0	-313.88	1.909	-----	0	-556.56	2.620	-----
Frank	0.097	-368.61	5.528	-----	0.093	-644.07	**7.112**	**0**
Gumbel	**0.24**	**-375.09**	**1.954**	**0.574**	0.032	-580.55	2.310	-----

According to the p-values of tests, we conclude that Gumbel copula and Frank copula have best fit to the two datasets respectively. Therefore Crime data are tail dependent, where TDC is equal to 0.574 and Stock data are tail independent. In the following, all proposed tests in Section 2 are performed on the two datasets and the results are displayed in Table 6. It should be noted that in carrying out these tests, for each dataset, the threshold c is chosen to have at least 30 observations greater than of the threshold value. Therefore, in two datasets, the thresholds are equal to –0.15 and –0.25 respectively.

Table 6. Independence tests for two datasets

Test	p.value	
	Crime Data	Stock Data
Neyman-Pearson	4.891685e-09	0.7543855
Fisher's κ	4.364887e-02	0.2194695
Kolmogorov-Smirnov	1.245545e-03	0.4993588
Chi-square	1.514254e-02	0.6754989
Cramer-von Mises	1.027966e-03	0.7278006
Anderson-Darling	3.082995e-04	0.6549564

In Crime data, all tests reject the null hypothesis of the tail independence of variables density and crmrte at 0.05 level, i.e., two variables density and crmrte are tail dependent; therefore, if the density of people per square mile exceeds a certain threshold, then crimes committed per person will exceed that specific threshold.

In Stock data, tail independence is not rejected by any of the tests at 0.05 level, i.e., stock prices of the two Japanese automakers Honda and Mazda are tail independent. Therefore tail independence tests confirmed the results of Table 5. It is noteworthy that if the TDC is estimated using the unsuitable copula function, the tail independence tests show this matter; this indicates the importance of using the test to verify the existence of tail dependence in the data.

5. Conclusion

In this paper, we recommended two new statistics Cramer-von Mises and Anderson-Darling for tail independence in extreme value models-based approach of Falk and Michel (2006). Simulations show that two tests are better than the proposed tests by Falk and Michel. Also, we illustrated the importance of using these tests by using two real datasets, while the tail dependence maybe is estimated incorrectly and this wrong is shown by tests.

References

Anderson, T.W., and Darling, D.A. 1954. "A Test of Goodness-of-Fit." *J. American Statistical Association* 49(268), 765-769.

Ane, T., and Kharoubi, C. 2003. "Dependence Structure and Risk Measure." *The journal of business*76(3), 411-438.

Coles, S., Heffernan, J., and Tawn, J. 1999. Dependence measures for extreme value analyses. *Extremes* 2(4), 339-365.

Cornwell, C. and Trumbull, W.N. 1994. "Estimating the Economic Model of Crime with Panel Data." *Review of Economics and Statistics* 76(2), 360-366.

Csorgo, S., and Faraway, J.J. 1996. "The Exact and Asymptotic Distributions of Cramer-von Mises Statistics." *Journal of the Royal Statistical Society. Series B (Methodological)* 58(1), 221-234.

Dupuis, D.J., and Tawn, J.A. 2001. "Effects of Mis–specification in Bivariate Extreme Value Problems." *Extremes* 4(4), 315-330.

Falk, M., and Michel, R. 2006. "Testing for Tail Independence in Extreme Value Models." *Annals of the Institute of Statistical Mathematics* 58(2), 261-290.

Frahm, G., Junker, M., and Schmidt, R. 2005. "Estimating the Tail-dependence Coefficient: Properties and Pitfalls." *Insurance: Mathematics and Economics* 37(1), 80-100.

Fuller, W.A. 1976. *Introduction to Statistical Time Series*. John Wiley, New York.

Geffroy, J., 1958. "Contribution a' La The'orie des Valeurs Extremes." *Publ. Inst. Statist. Univ. Paris* 7, 37-121.

Geffroy, J. 1959. A' La The'orie Des Valeurs Extremes II." *Publ. Inst. Statist. Univ. Paris* 8, 3-65.

Genest, C., Rémillard, B., and Beaudoin, D. 2009. "Goodness-of-fit Tests for Copulas: A Review and a Power Study." *Insurance: Mathematics and Economics* 44(2), 199-213.

Joe, H., 2014. *Dependence Modeling with Copulas*. CRC Press.

Malevergne Y., Sornette D. 2004. *Investigating Extreme Dependencies. Extreme Financial Risks*. (From dependence to risk management), Springer, Heidelberg.

Nelsen, R.B. 2006. *An Introduction to Copulas*. Springer, New York.

Pickands III, J., 1981. "Multivariate Extreme Value Distributions." In *Proceedings of the 43th Session of the International Statistical Institute (Buenos Aires)* 859-878.

Sibuya, M. 1960. "Bivariate Extreme Statistics." *Annals of the Institute of Statistical Mathematics* 11(2), 195-210.

Sklar, A. 1959. "Fonctions de Re'partition a' n Dimensions et Leurs Marges." *Publ. Inst. Statist. Univ. Paris* 8, 229-231.

Guidelines for Authors

The Philippine Statistician (TPS) is the official scientific journal of the Philippine Statistical Association. Inc. (PSA). It considers papers resulting from original research in statistics and its applications. Papers will be sent for review on the assumption that this has not been published elsewhere nor is submitted in another journal.

Aims and Scope

The Journal aims to provide a media for the dissemination of research by statisticians and researchers using statistical method in resolving their research problems. While a broad spectrum of topics will be entertained, those with original contribution to the statistical science or those that illustrates novel applications of statistics in solving real-life problems will be prioritized. The scope includes, but is not limited to the following topics:

- Official Statistics
- Simulation Studies
- Survey Sampling
- Time Series Analysis
- Nonparametric Methods
- Econometric Theory and Applications
- Other Applications
- Computational Statistics
- Mathematical Statistics
- Statistics Education
- Biostatistics
- Experimental Designs and Analysis

In addition to research articles, the Journal will have the following sections that may appear in some of its issues (but not necessarily in all):

- *Letters to the Editor*. This section will provide a forum for the airing of opinions on issues pertinent to the statistical community or offers commentaries on articles that have appeared in the journal.
- *Notes* section will include notices and announcements of upcoming events, conferences, calls for papers.
- *Review* section will present reviews on statistics books and software.
- *Teacher's Corner* shares experience of some teachers related to statistical education.

Articles submitted for the four special sections above will be reviewed only by the Editor and/or Associate Editors.

Submission of Manuscript

Only unpublished manuscripts will be considered. They will be refereed and evaluated on content, language and presentation. The article in MS word format, without author's identification should be sent as email attachment to Jose Ramon Albert, Editor, The Philippine Statistician: jrgalbert@gmail.com. A separate file containing the title of the paper, authors(s) (corresponding author identified) and their affiliations and complete address should be included in a separate file to be emailed along with the main article. To avoid delays and difficulties in submission, authors should follow instructions on style prescribed below.

A printed page in the Journal will have a maximum amount of space of 4.5" by 8.5."

Organization of Manuscript

The manuscript should be written in 8.5"x11" page using Times New Roman font size 12 with 1" margin in all sides.

Manuscripts must be organized in the following manner:

- *Title*: This should be brief and concise.
- *Abstract and Key Words*: An abstract of at most 250 words must be submitted with the manuscript. It precedes the article text. The abstract should summarize objectives, results, and main conclusions, but it should not contain any graph or complex mathematical notation and no references. Three to six keywords should be identified.
- *Article Text*: Sections should be concise and numbered in a decimal system. Tables, Figures, and Artwork may be used within the body of the article, should be numbered consecutively. Tables, Figure Titles and Legends, Figure Artwork must be strategically placed in the article text. The original files for the tables, figures, and artwork should also be submitted to facilitate typesetting. Authors must obtain written permission to reproduce or adapt all or part of a figure from a copyrighted source. Mathematical equations cited in the text should be numbered consecutively. Numbers should be placed at the rightmost margin of the equation line in parenthesis. Matrices should appear in bold and vectors in italics. All other symbols should appear in italics. The preferred software for equations are Mathtype and MS Equation Editor, which are add-ins of MS Word.
- *Acknowledgments*: An acknowledgment section may be included at the end of the article. This section should acknowledge financial assistance in the form of grants or university funding, assistance by individual colleagues, and any other pertinent information. This section will be inserted by the author only upon acceptance of the paper.
- *References*: All references included in the list at the end of an article must be cited in the text. References are cited in the text in the following format: (Author Year). Up to two authors can be cited in the text. If there are three or more authors, only the first author will be cited in the text, e.g. (Author et al. Year). The following format will be followed in the listing references:

<u>Journal Article</u>

LANDAGAN, O. and BARRIOS, E. 2007. "An Estimation Procedure for a Spatial-temporal Model." *Statistics and Probability Letter* 77(4): 401-406.

Book

KOTTAK, C. 1991. *Anthropology: The Exploration of Human Diversity*, 5th ed., New York: McGraw-Hill, Inc.

Book Chapter

FINK, E. and PRATT, K. 2008. "Indexing Compressed Time Series." In Last, M., Kandel, A. and Bunke, H., eds., *Data Mining in Time Series Databases*, Singapore: World Scientific, pp. 43-66.

Internet Document

MUNDLAK, Y., LARSON, D. and BUTZER, R. 2002. "Determinants of agricultural growth in Indonesia, the Philippines, and Thailand, V. 1." World Bank Working Paper 2803, The World Bank. Available at: http:// econ. worldbank.org/external/default/main?pagePK=64165259&piPK= 6 4 1 6 5 4 2 1&theSitePK=469372&menuPK=64216926&entityID= 000094946_02032604542948

For particulars about the style, please download the Philippine Statistician Style Guide at www.philstat.org.ph.

- *Appendices*: A single appendix is headed, “APPENDIX: FOLLOWED BY A DESCRIPTIVE TITLE”. If there are two or more appendices, they should be labeled, “APPENDIX A,” “APPENDIX B,” and so on.
- *Editorial Style*: In addition to content, manuscripts are evaluated on their conciseness and clarity. Thus, the Journal gives premium to well-written and well-structured papers that will be of interest to a wide segment of the readership. Manuscripts and reviews that have been accepted for publication will be copy edited in accordance to accepted rules of correct grammar, usage, spelling, and punctuation. To avoid common problems of style, for guidelines on style, usage, and the preparation of technical manuscripts for publication, the following reference may be consulted:

  The Chicago Manual of Style (14th ed.) (1993),
  Chicago: University of Chicago Press.

Editorial Notes

Use quotation marks only when a standard term is used in a nonstandard way and to indicate the beginning and ending of a direct quotation.

1. Hyphens are used when two or more adjectives or an adjective and a noun together modify another noun; for example, *goodness-of-fit test* is the equivalent of *test for goodness of fit*. Most words with prefixes such as sub and non are not hyphenated, for example, *subtable, nonnormal*.
2. Italics are used to introduce important terms, when appropriate; they are to be used sparingly to indicate emphasis.

3. Abbreviations and acronyms should be minimized; those that are used are spelled out on their first appearances in the manuscript with the shortened form given in parentheses, for example, best linear unbiased estimate (BLUE).
4. Numbers under 10 are spelled out when they are not part of an equation or an expression containing symbols.
5. The sign % is always used when giving a specific percentage, for example, 23%, not 23 percent. Otherwise use the word percent.

Copyright Transfer Form

Authors of accepted papers will be required to submit an author copyright transfer form before the final release of the journal.

PSAI OFFICERS AND BOARD OF DIRECTORS 2021

EXECUTIVE COMMITTEE MEMBERS

President	Dennis S. Mapa
Vice President	Carmelita N. Ericta
Secretary	Benjamin Arsenio Y. Navarro
Treasurer	Jade Eric T. Redoblado
Immediate Past President	Lisa Grace S. Bersales

PSAI BOARD OF DIRECTORS

A. Individual Members

Jose Ramon G. Albert
Joselito R. Basilio
Rosalinda P. Bautista
Teresita B. Deveza
Carmelita N. Ericta
Dennis S. Mapa
Romeo S. Recide
Jade Eric T. Redoblado

B. Institutional Members - Government Sector

Philippine Statistics Authority	*Represented by*	Benjamin Arsenio Y. Navarro
Philippine Statistical Research and Training Institute		Josefina V. Almeda
University of the Philippines School of Statistics		Joseph Ryan G. Lansangan

C. Institutional Members - Private Sector

Ateneo de Zamboanga University	Jocelyn Partosa
Ateneo Social Science Research Center	Frances Michelle C. Nubla
Social Weather Stations	Gerardo A. Sandoval

D. PSAI Regional Chapter

PSAI Region 8 Chapter	Wilma A. Perante

PSAI Working Committees 2021

Advocacy Committee
Chair: Dennis S. Mapa
Co-Chair: Lisa Grace S. Bersales

Annual Conference Committee
Chair: Carmelita N. Ericta
Co-Chair: Francisca N. Dayrit

Sub-Committee: Scientific Program
Chair: Joseph Ryan G. Lansangan

Annual Meeting and Christmas Party Committee
Chair: Gian Louisse Roy
Co-Chair: Mi-Auree L. Bautista

Finance Committee
Chair: Jade Eric T. Redoblado
Co-Chair: Josefina V. Almeda

Institutional Development Committee
Chair: Romeo S. Recide
Co-Chairs: Lisa Grace S. Bersales
Teresita B. Deveza

Institutional Training Committee
Chair: Josefina V. Almeda
Co-Chair: Joseph Ryan G. Lansangan

Membership Committee
Chair: Ferdinand S. Co
Co-Chair: Rosalinda P. Bautista

Nominations and Election Committee
Chair: Ludivinia D. Gador
(January to March)
Chair: Estela T. de Guzman
(April to December)

Publications Committee
Chair: Jose Ramon G. Albert
Jose Ramon G. Albert (Editor, TPS)
Jana Flor V. Vizmanos (Managing Editor, TPS)
Mika S. Muñoz (Editorial Coordinator, TPS)
Genelyn F. Sarte (Editor, PSAI Newsletter)

Search and Awards Committee
Chair: Luisito Asuncion
Co-Chair: Maria Praxedes R. Peña

Social Media, Information and Communications Committee
Chair: Benjamin Arsenio Y. Navarro
Co-Chair: Gerardo A. Sandoval

Ad Hoc Committee on Regional Affairs
Chair: Wilma A. Perante
Co-Chairs: Jocelyn D. Partosa
Frances Michelle C. Nubla

Web Development Team
Lead Web Developer/
Administrator: Ferdinand S. Co