Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v12n2p508

A particle swarm optimization method for variable selection in beta regression model
By Algamal

Published: 14 October 2019

Electronic Journal of Applied Statistical Analysis
Vol. 12, Issue 02, October 2019, 508-519
DOI: 10.1285/i20705948v12n2p508

A particle swarm optimization method for variable selection in beta regression model

Zakariya Yahya Algamal*

Department of Statistics and Informatics, University of Mosul

Published: 14 October 2019

Beta regression model has received much attention in several science fields in modeling proportions or rates data. Selecting a small subset of relevant variables from a large number of variables is an important task for building a predictive regression model. This paper proposes employing the particle swarm optimization algorithm as a variable selection method in the beta regression model with varying dispersion. The performance of the proposed method is evaluated through simulation and real data application. Results demonstrate the superiority of the proposed method compared to other competitor methods including corrected Akaike information criterion, corrected Schwarz information criterion, and corrected Hannan and Quinn criterion. Thus, the proposed method can efficiently helpful as a variable selection tool in the beta regression model with varying dispersion.

keywords: Variable selection; beta regression model; varying dispersion; particle swarm optimization algorithm.

1 Introduction

In regression modeling, the response variable can be a continuous variable in form of proportions or rates, such as the fraction of income contributed to a retirement fund and the percentage of ammonia escaping unconverted from an oxidation plant (Ospina and Ferrari, 2012), where the values are limited to the interval (0, 1) (Branscum et al., 2007). The classical linear regression which is based on the ordinary least square method

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

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

is inappropriate for such situations (Ospina and Ferrari, 2012). Consequently, Ferrari and Cribari-Neto (2004) introduced beta regression model in which the response variable is distributed from the beta distribution.

The increasing trend of measuring and collecting a large number of variables becomes popular in many real applications. However, these datasets are often containing a large number of redundant and irrelevant variables that may significantly degrade the model prediction accuracy. In the regression modeling, the existence of large numbers of variables can degrade the regression model. As a result, selecting a small subset of relevant variables from a large number of variables is an important task for building predictive regression models.

Searching for the best subset of variables is known to be an NP-hard problem where it requires a long time for computing associated with high cost. The traditional variable selection methods, such as stepwise selection, information criteria, and backward elimination computationally become more expensive. In recent years, the meta-heuristics algorithms, such as genetic algorithm, ant colony optimization algorithm, particle swarm optimization algorithm, and crow search algorithm, are widely applied as variable selection methods. This is because that the variable selection is considered as an optimization problem in which it minimizes the number of selected variables while maintaining the maximum accuracy of prediction (Qasim et al., 2018; Algamal et al., 2017; Alanaza and Algamal, 2018; Algamal, 2017).

The main target in this paper is to propose the particle swarm optimization algorithm, which is a swarm intelligence approach, as a variable selection method in the beta regression model. The proposed algorithm would efficiently help in finding the most important variables in the beta regression model with a high prediction. The advantage of the proposed algorithm is proved through simulation study and a real data application. The remainder of our paper is organized as follows. Sections 2 and 3 cover the description of the beta regression model and the variable selection methods. The details of the particle swarm optimization algorithm are illustrated in Section 4. The expression of the proposed method is explained in Section 5. Section 6 is devoted to simulation and real data application results. The conclusion is covered in Section 7.

2 Beta regression model

In beta regression model (BRM), the response variable, y, is assumed to follow beta distribution. The probability density function of beta distribution is given by

$$f\left(y;\theta_1,\theta_2\right)=\frac{\Gamma(\theta_2)}{\Gamma(\theta_1\theta_2)\Gamma((1-\theta_1)\theta_2)}(y)^{\theta_1\theta_2-1}(1-y)^{((1-\theta_1)\theta_2)-1},\quad 0<y<1, \qquad (1)$$

where $0<\theta_1<1$ and $\theta_2>0$. The mean and variance of Eq. (1) are given by, respectively, $E(y)=\theta_1$ and $V(y)=\theta_1(1-\theta_1)/(1+\theta_2)$ where θ_2 is a dispersion parameter. For a fixed value of θ_1, the $V(y)$ value decrease when the value of θ_2 increases.

Consider that we have a data set $\{(y_i,\mathbf{x}_i)\}_{i=1}^n$ where $y_i\in R$ is a response variable belongs to Eq. (1), $\mathbf{x}_i=(x_{i1},x_{i2},...,x_{ip})\in R^p$ is a $p\times 1$ known explanatory variable

vector, then in BRM, the mean is related to the explanatory variables as

$$g(\theta_{1i}) = \mathbf{x}_i{}^T\beta = \eta_i, \tag{2}$$

where $\beta = (\beta_0, \beta_1, ..., \beta_p)$ is a $(p+1)\times 1$ vector of unknown regression coefficients. Logit, probit, cloglog, and loglog are the used link functions of Eq. (2).

Ferrari and Cribari-Neto (2004) extended the BRM to allow θ_2 to vary across observations. The BRM with varying dispersion (BRMVD) is defined as

$$\begin{aligned} g(\theta_{1i}) &= \mathbf{x}_i{}^T\beta = \eta_i \\ \mathbf{h}(\theta_{2i}) &= \mathbf{s}_i{}^T\alpha = \vartheta_i, \end{aligned} \tag{3}$$

where $\alpha = (\alpha_1, ..., \alpha_k)$ is a $k \times 1$. vector of unknown regression coefficients and $\mathbf{s}_i = (s_{i1}, s_{i2}, ..., s_{ik}) \in R^k$ is a $k \times 1$ known explanatory variable vector in addition to $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})$which are not exclusive, $p + k < n$.

The log-likelihood function of Eq. (3) is given by

$$\begin{aligned} \ell(\beta, \alpha) &= \sum_{i=1}^{n} \ell_i(\theta_{1i}, \theta_{2i}) \\ &= \ln\Gamma(\theta_{2i}) - ln\Gamma((1-\theta_{1i})\,\theta_{2i}) + (\theta_{1i}\,\theta_{2i} - 1)lny_i \\ &\quad +\{((1-\theta_{1i})\,\theta_{2i}) - 1\}\ln(1-y_i)\,, \end{aligned} \tag{4}$$

where $\theta_{1i} = g^{-1}(\eta_i)$ and $\theta_{2i} = \mathbf{h}^{-1}(\vartheta_i)$. Differentiation of Eq. (4) with respect to the β and α , respectively, is defined as

$$\mathbf{U}_\beta(\beta, \alpha) = \frac{\partial\ell(\beta, \alpha)}{\partial\beta} = \sum_{i=1}^{n} \theta_{2i}(\tilde{y}_i - \tilde{\theta}_{1i})\,\frac{\mathbf{d}\theta_{1i}}{\mathbf{d}\eta_i}\frac{\partial\eta_i}{\partial\beta_p}, \tag{5}$$

$$\mathbf{U}_\alpha(\beta, \alpha) = \frac{\partial\ell(\beta, \alpha)}{\partial\alpha} = \sum_{i=1}^{n} \left\{ \begin{array}{l} \theta_{1i}(\tilde{y}_i - \tilde{\theta}_{1i}) + \psi(\theta_{2i}) \; - \psi(1-\theta_{1i})\,\theta_{2i} \\ \quad + \ln(1-y_i) \end{array} \right\} \frac{\mathbf{d}\theta_{2i}}{\mathbf{d}\vartheta_i}\frac{\partial\vartheta_i}{\partial\alpha_k}, \tag{6}$$

where $\tilde{y}_i = \ln(y_i/(1-y_i))$, $\tilde{\theta}_{1i} = \psi(\theta_{1i}\theta_{2i}) - \psi((1-\theta_{1i})\,\theta_{2i})$, $\psi(.)$ represents the digamma function, $\mathbf{d}\theta_{1i}/\mathbf{d}\eta_i = 1/g'(\theta_{1i})$, and $\mathbf{d}\theta_{2i}/\mathbf{d}\vartheta_i = 1/h'(\theta_{2i})$. Then the maximum likelihood estimator of β and α are obtained from the solution of the nonlinear system $\mathbf{U}(\xi) = 0$, where $\xi = (\beta^T, \alpha^T)^T$ (Simas et al., 2010).

3 Variable selection for the BRM

Variable selection procedure has been widely used in many applications. Usually, some variables may be redundant and others may be irrelevant. As a result, these extra variables can increase computational time and can have a negative impact on the prediction accuracy. Therefore, selecting the most relevant variables out of the whole variables leads to increase the prediction accuracy and to simplify the model interpretation.

In the literature, Zhao et al. (2014); Qasim (2019) introduced variable selection in BRMVD by proposing the penalized method, which includes the least absolute shrinkage and selection operator (LASSO), the smoothly clipped absolute deviation (SCAD), and the minimax concave penalty (MCP). The estimation of ξ depending on the penalized likelihood function is given by

$$\hat{\xi}_{\textbf{penalized}} = \arg\max_{\xi} \left[\ell(\beta, \alpha) - n \sum_{j=1}^{p} P_{\lambda_1}(|\beta_j|) - n \sum_{l=1}^{k} P_{\lambda_2}(|\alpha_j|) \right]. \tag{7}$$

Further, Bayer and Cribari-Neto (2015a) introduced several model selection criteria in BRMVD. Besides, they proposed a fast two-step model selection scheme. On the other hand, Bayer and Cribari-Neto (2015b) proposed a bootstrap-based model selection criteria in BRMVD. They introduced two new selection criteria. The first one is the bootstrapped likelihood quasi-cross validation (CV), while the second one is its 632QCV variant.

4 Particle swarm optimization algorithm

In recent years, population-based swarm intelligence algorithms, which are a class of natural-inspired algorithms, are widely used for solving complex optimization problems (Lin et al., 2008; Algamal, 2019). Particle swarm optimization (PSO) is one of the most powerful algorithms because of its easiness in implementation with few parameters (Xia et al., 2017). The PSO algorithm is originally introduced by Eberhart and Kennedy (1995) inspiring from the social behavior associated with fish schooling and bird flocking.

In PSO, the swarm contains a number of particles, where each particle is considered as an individual. In addition, the solution space of the optimization problem is stated as a search space in which each particle is considered as a solution to the problem. All particles move according to their velocities through a d-dimension search space. At each iteration of the algorithm, the movement of each particle is calculated as follows:

$$z_i(t+1) = z_i(t) + v_i(t+1), \tag{8}$$

$$v_i(t+1) = w \times v_i(t) + a_1 \times b_1 \times (Pbest_i(t) - z_i(t)) + a_2 \times b_2 \times (Gbest_i(t) - z_i(t)), \tag{9}$$

where $v_i(t)$ and $z_i(t)$, respectively, is the velocity and the position of particle i at iteration t, $Pbest_i(t)$ is the best position that is found by particle i, and $Gbest_i(t)$ is the best position that is found by swarm as a whole. Further, $w$is the inertia weight, a_1 and a_2 are the acceleration coefficients. While, b_1 and b_2 are random values selected from a uniform distribution within the range of 0 and 1. For a swarm consists of m particles and the objective function, f, is used to calculate the fitness of the particles with a maximization or minimization task, the personal best values, and the global best value are updated at iteration t, respectively, as follows:

$$Pbest_i(t+1) = \begin{cases} Pbest_i(t) & \textbf{if} \quad f(Pbest(t)) \leq f(z_i(t+1)) \\ z_i(t+1) & \textbf{if} \quad f(Pbest(t)) > f(z_i(t+1)), \end{cases} \tag{10}$$

$$Gbest_i(t+1) = \min(max)\{ f(h),\ f(Gbest_i(t))\}, \tag{11}$$

where $h \in \{Pbest_1(t), ..., Pbest_m(t)\}$.

5 The proposed method

Originally, PSO is proposed to solve the continuous optimization problems. However, for performing the variable selection, the optimization problem is not continuous. A binary PSO (BPSO) is adapted to perform variable selection. In contrast to PSO, the position in BPSO is binary in which the value 1 represents that the variable is important and 0 otherwise. In other words, if the i^{th} variable is included in the model, then $\mathbf{x}_i = 1$, otherwise, $\mathbf{x}_i = 0$. In variable selection, the dimension of each particle is the number of the original variables in the model. In BPSO, the velocity of each particle is needed to transfer into a probability vector because the position is binary. A common transfer function is a sigmoid function (sigm) which is defined as

$$\mathbf{sigm}_i = \frac{1}{1+\exp[-v_i(t)]}. \tag{12}$$

Consequently, the position in Eq. (8) is updated as follows:

$$z_i(t+1) = \begin{cases} 1 & \textbf{if } b_3 < \mathbf{sigm}_i \\ 0 & \textbf{otherwise}, \end{cases} \tag{13}$$

where b_3 is a random number generated from the uniform distribution between 0 and 1.

Accordingly, our proposed algorithm setting for performing variable selection in BR-MVD is as follows:

1. The number of particles, m, is set to 50 and the maximum number of iterations is $\mathbf{t}_{\max}$= **100**. The acceleration coefficients a_1 and a_2 are set within the range [1.5, 4]. The a_1 and a_2 are updating during the iteration as following:

$$a_1 = a_{1,\min} + \frac{t}{t_{\max}}(a_{1,\max} - a_{1,\min}), \tag{14}$$

$$a_2 = a_{2,\min} + \frac{t}{t_{\max}}(a_{2,\max} - a_{2,\min}). \tag{15}$$

   Besides, the $w$is set with minimum and maximum values as: $w_{\min} = 0.1$ and $w_{\max} = 0.99$, and it is updating as:

$$w = w_{\max} - \frac{t}{t_{\max}}(w_{\max} - w_{\min}). \tag{16}$$

2. The positions of each particle are randomly specified from the uniform distribution with 0 and 1. Here, the positions are represented by the explanatory variables of each $\mathbf{s}_i = (s_{i1}, s_{i2}, ..., s_{ik})$ and $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})$. The representation of the positions of a particle is explained in Figure 1.

3. The initial velocity of each particle is generated from a uniform distribution within the range [0, 6].

4. The fitness function, f, is defined as

$$f = \min\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right], \tag{17}$$

   where each particle has a fitness value, and, therefore, the personal best values and the global best value are calculated.

5. The velocities and the positions of the particles are updated using Eq. (9) and Eq. (13), respectively.

6. Steps 4 and 5 are repeated until a $t_{\max}$ is reached.

x_1	x_2	.....	x_{p-1}	x_p	s_1	s_2	.....	s_{k-1}	s_k
1	0	.....	1	0	1	1	.....	0	1

Figure 1: The representation of the particle position.

6 Computational results

In this section, the performance of our proposed method, PSO- BRMVD is tested. Further, the performance of PSO- BRMVD is compared with other variable selection methods that were used in Bayer and Cribari-Neto (2015a). They are:

1. The corrected Akaike information criterion (CAIC)

$$\mathbf{CAIC} = -2\ell(\hat{\beta}, \hat{\alpha}) + \frac{2n\,(p+k)}{n-(p+k)-1}. \tag{18}$$

2. The corrected Schwarz information criterion (CSIC)

$$\mathbf{CSIC} = -2\ell(\hat{\beta}, \hat{\alpha}) + \frac{n\,(p+k)ln(n)}{n-(p+k)-1}. \tag{19}$$

3. The corrected Hannan and Quinn criterion (CHQ)

$$\mathbf{CHQ} = -2\ell(\hat{\beta}, \hat{\alpha}) + \frac{2n\,(p+k)ln(ln(n))}{n-(p+k)-1}. \tag{20}$$

6.1 Monte Carlo simulation study results

In this section, the performance of PSO- BRMVD is evaluated. The sample size is considered with $n \in \{30, 50, 100, 200\}$ and the response variable is generated from the following distribution

$$y_i \sim \mathbf{beta}(\theta_{1i}\,\theta_{2i}, (1-\theta_{1i})\,\theta_{2i}), \tag{21}$$

where θ_{1i} and θ_{2i} are generated according to the logit link function as

$$\begin{aligned} \theta_{1i} &= \frac{\exp(\mathbf{x}_i{}^T\beta)}{1+\exp(\mathbf{x}_i{}^T\beta)}, \\ \theta_{2i} &= \frac{\exp(\mathbf{s}_i{}^T\alpha)}{1+\exp(\mathbf{s}_i{}^T\alpha)}, \end{aligned} \tag{22}$$

where the variables $\mathbf{x}_i$ and $\mathbf{s}_i$ are generated from the uniform distribution with 0 and 1. The true parameter vector

$$\beta$$

and α are set as $\beta = (1, 1, -0.5, 1.5, \underbrace{0, ..., 0}_{p-4})^T$ and $\alpha = (1, 1, -1.5, 0.5, \underbrace{0, ..., 0}_{k-4})^T$ with $\beta_0 = \alpha_0 = 0$. In each sub-model, there are 4 important variables and the rest is irrelevant variables. In this situation, two cases are considered:

Case 1: In this case $p = k = 8$.

Case 2: In this case $p = k = 15$.

The performance of the PSO-BRMVD is assessed by the following four criteria: (1) the mean squared error (MSE) as $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2/n$; the number of the true zero coefficients correctly identified as zeros (TZ); the number of the truly nonzero coefficients incorrectly identified as zeros (INZ); the percentage of correctly estimated BRMVD (PC). The higher the values of PC and TZ, and the lower the values of MSE and INZ, the better the variable selection performance is. All the computations of our paper were conducted using R. Depending on 500 times of generated the data, the averaged MSE, TZ, INZ, and PC with their associated standard deviations (the number in parentheses) are reported in Tables 1 and 2, respectively, for case 1 and case 2. From these tables, some general remarks are made. First, it is seen that the PSO- BRMVD is able to achieve lower MSE than the CAIC, CHQ, and the CSIC for all cases. Meanwhile, CAIC presents a large MSE among all methods. For instance, in Table 2 when $n = 30$, the MSE reduction by PSO- BRMVD was about 43.87%, 35.30%, and 31.76% comparing with CAIC, CHQ, and CSIC, respectively. Further, the PSO- BRMVD is always showing the smallest MSE among the competitor methods regardless of the value of . In general, the performance of CAIC, CHQ, and the CSIC, in terms of MSE, tends to improve with an increasing sample size. Second, in terms of TZ criterion, the variable selection results obtained by the PSO- BRMVD are obviously closed to the true nonzero coefficients for both the mean sub-model and dispersion sub-model. On other words, PSO- BRMVD selected, on average, more than 7 important variables out of 8 true variables in case 1, while, in case 2, PSO- BRMVD selected, on average, more than 20 important variables out of 22 true

variables. This indicating that the PSO- BRMVD is the best comparing with CAIC, CHQ, and the CSIC. For example, in Table 2 when , PSO- BRMVD truly selects, on average, nearly about 21 relevant variables out of 22 important variables. While CAIC, CHQ, and CSIC select no more than 18 relevant variables. Third, in terms of INZ criterion, there is an obvious trend that PSO- BRMVD selects a very few unimportant variables of the mean sub-model and dispersion sub-model comparing with CAIC, CHQ, and CBIC, for both cases, where the number of the true nonzero coefficients, on average, which are correctly set to zero is low compared with others. In conclusion, it is obvious that the simulation results for the BRMVD demonstrated the superior using of PSO-BRMVD in variable selection. Besides, it is concluded from the simulation results that the PSO- BRMVD performance in variable selection is not changed by changing the number of true zero coefficients and the sample size.

Table 1: Case 1 results, on average, for BRMVD

Methods	MSE	TZ	INZ	PC
	$n = 30$			
PSO- BRMVD	3.221 (0.011)	7.547 (0.012)	0.414 (0.010)	0.911 (0.005)
CAIC	7.061 (0.030)	4.132 (0.017)	3.241 (0.018)	0.678 (0.010)
CHQ	6.455 (0.022)	5.022 (0.018)	2.872 (0.018)	0.779 (0.008)
CSIC	5.811 (0.019)	5.971 (0.018)	2.066 (0.012)	0.836 (0.008)
	$n = 50$			
PSO- BRMVD	3.102 (0.011)	7.577 (0.013)	0.220 (0.010)	0.934 (0.006)
CAIC	7.574 (0.031)	4.426 (0.022)	3.197 (0.025)	0.671 (0.009)
CHQ	6.285 (0.019)	5.135 (0.019)	2.928 (0.022)	0.785 (0.007)
CSIC	5.551 (0.019)	5.691 (0.019)	2.132 (0.012)	0.854 (0.006)
	$n = 100$			
PSO- BRMVD	3.065 (0.013)	7.631 (0.013)	0.295 (0.012)	0.938 (0.005)
CAIC	7.425 (0.033)	4.493 (0.019)	3.172 (0.023)	0.680 (0.008)
CHQ	6.246 (0.021)	5.232 (0.019)	3.503 (0.021)	0.791 (0.006)
CSIC	5.705 (0.023)	6.035 (0.016)	1.712 (0.017)	0.862 (0.005)
	$n = 200$			
PSO- BRMVD	2.115 (0.013)	7.648 (0.013)	0.215 (0.012)	0.947 (0.004)
CAIC	7.134 (0.033)	4.423 (0.019)	3.172 (0.023)	0.685 (0.008)
CHQ	6.107 (0.021)	5.232 (0.019)	2.513 (0.021)	0.798 (0.006)
CSIC	5.271 (0.023)	6.015 (0.016)	1.815 (0.017)	0.867 (0.005)

Table 2: Case 2 results, on average, for BRMVD

Methods	MSE	TZ	INZ	PC
	$n = 30$			
PSO- BRMVD	5.471 (0.013)	21.732 (0.014)	0.014 (0.011)	0.925 (0.006)
CAIC	9.748 (0.021)	14.528 (0.022)	3.106 (0.018)	0.772 (0.009)
CHQ	8.456 (0.021)	16.543 (0.021)	2.772 (0.018)	0.783 (0.008)
CSIC	8.018 (0.018)	17.118 (0.017)	1.086 (0.012)	0.841 (0.006)
	$n = 50$			
PSO- BRMVD	5.104 (0.013)	20.951 (0.015)	0.125 (0.010)	0.917 (0.007)
CAIC	9.364 (0.021)	15.592 (0.023)	3.066 (0.024)	0.767 (0.010)
CHQ	8.275 (0.021)	16.805 (0.021)	2.525 (0.022)	0.788 (0.009)
CSIC	7.747 (0.017)	18.176 (0.019)	1.112 (0.012)	0.861 (0.008)
	$n = 100$			
PSO- BRMVD	5.029 (0.014)	21.311 (0.015)	0.088 (0.012)	0.933 (0.005)
CAIC	9.389 (0.022)	15.601 (0.021)	3.204 (0.022)	0.780 (0.009)
CHQ	8.212 (0.019)	16.816 (0.021)	2.638 (0.021)	0.792 (0.007)
CSIC	7.669 (0.019)	18.887 (0.021)	1.414 (0.015)	0.873 (0.007)
	$n = 200$			
PSO- BRMVD	4.885 (0.011)	21.911 (0.014)	0.081 (0.012)	0.941 (0.005)
CAIC	9.044 (0.028)	14.601 (0.024)	3.573 (0.022)	0.785 (0.008)
CHQ	7.937 (0.022)	15.816 (0.022)	2.623 (0.021)	0.804 (0.006)
CSIC	7.118 (0.021)	18.187 (0.021)	1.614 (0.017)	0.882 (0.005)

6.2 real application results

In this section, a real application is considered for testing our proposed method to a data from a body fat study, which had been analyzed by Zhao et al. (2014). In this data, there are 252 observations for body fat patients on 13 explanatory variables, of which the y is a quantitative measurement of the percentage of body fat. The 13 explanatory variables include age (years) (x_1); weight (pounds) (x_2); height (inches) (x_3); neck circumference (cm) (x_4); chest circumference (cm) (x_5); abdomen circumference (cm) (x_6); hip circumference (cm) ((x_7); thigh circumference (cm) (x_8); knee circumference (cm) (x_9); ankle circumference (cm) (x_{10}); extended biceps circumference (x_{11}); forearm circumference (cm) (x_{12}) and wrist circumference (cm) (x_{13}). Related to citez30, the response variable is following the beta distribution, and, thus the BRMVD is been more suitable regression model with the logit link function to the mean sub-model and the identity link function for the dispersion sub-model. Depending on the BRMVD analysis, after exclusion the three identified outlier observations, five explanatory variables, x_2, x_6, x_7, x_{12}, and x_{13}, are significantly related to the response variables with a level

of significant 0.05 for the mean sub-model, and one variable, s_2, is significant. The performance results of the used methods are summarized in Table 3.

It is seen from the result of Table 3 that PSO- BRMVD clearly succeeds to select the most significant variables for both mean and dispersion sub-models except x_6. On the other hand, the PSO- BRMVD is able to produce the better prediction accuracy by reducing the MSE comparing with CAIC, CHQ, and CSIC. Furthermore, it is obvious that CAIC, CHQ, and CSIC selected non-statistically significant variables. For example, CAIC selected x_4., x_9, s_4, and s_{10} which they are not significant. Among the significant variables, variables, x_2, x_7, s_2 are the three common important selected variables by the used methods.

Table 3: The selected variables and MSE body fat data

Methods	Selected variables		MSE
	Mean sub-model	Dispersion sub-model	
PSO- BRMVD	x_2, x_7, x_{12}, x_{13}	s_2	251.731
CAIC	x_2, x_4, x_7, x_9, x_{12}	s_2, s_4, s_{10}	463.109
CHQ	x_2, x_4, x_7, x_{12}	s_2, s_4	418.749
CSIC	x_2, x_4,x_7 , x_{13}	s_2, s_{10}	362.114

7 Conclusion

In this work, the problem of selecting variables in varying dispersion beta regression model is investigated. A particle swarm optimization algorithm was proposed to perform the variable selection. Simulation and real data application are carried out not only for the PSO-BRMVD but also for other alternative methods. The obtained results prove the dominance of the PSO-BRMVD against CAIC, CHQ, and CSIC in terms of MSE, TZ, INZ, and PC.

8 Acknowledgment

The author is very grateful to the University of Mosul/ College of Computer Sciences and Mathematics for their provided facilities, which helped to improve the quality of this work.

References

Alanaza, M. M. and Algamal, Z. Y. (2018). Proposed methods in estimating the ridge regression parameter in poisson regression model. *Electronic Journal of Applied Statistical Analysis*, 11:506–515.

Algamal, Z. (2017). An efficient gene selection method for highdimensional microarray data based on sparse logistic regression. *Electronic Journal of Applied Statistical Analysis*, 10:242–256.

Algamal, Z. (2019). Variable selection in count data regression model based on firefly algorithm. *STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING*, 7:520–529.

Algamal, Z. Y., Qasim, M. K., and Ali, H. T. (2017). A qsar classification model for neuraminidase inhibitors of influenza a viruses (h1n1) based on weighted penalized support vector machine. *SAR and QSAR in Environmental Research*, 28:415?426.

Bayer, F. M. and Cribari-Neto, F. (2015a). Bootstrap-based model selection criteria for beta regressions. *Test*, 24(4):776–795.

Bayer, F. M. and Cribari-Neto, F. (2015b). Model selection criteria in beta regression with varying dispersion. *Communications in Statistics - Simulation and Computation*, 46(1):729–746.

Branscum, A. J., Johnson, W. O., and Thurmond, M. C. (2007). Bayesian beta regression: Applications to household expenditure data and genetic distance between foot-and-mouth disease viruses. *Australian & New Zealand Journal of Statistics*, 49(3):287–301.

Eberhart, R. and Kennedy, J. (1995). A new optimizer using particle swarm theory. In *Micro Machine and Human Science, 1995. MHS'95., Proceedings of the Sixth International Symposium on*, pages 39–43. IEEE.

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815.

Lin, S.-W., Ying, K.-C., Chen, S.-C., and Lee, Z.-J. (2008). Particle swarm optimization for parameter determination and feature selection of support vector machines. *Expert Systems with Applications*, 35(4):1817–1824.

Ospina, R. and Ferrari, S. L. P. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, 56(6):1609–1623.

Qasim, M. K. (2019). Modified nanostructure mgo superbasicity with cao in heterogeneous transesterification of sunflower oil. *Egyptian Journal of Chemistry*, 36:475–485.

Qasim, M. K., Algamal, Z. Y., and Ali, H. T. (2018). A binary qsar model for classifying neuraminidase inhibitors of influenza a viruses (h1n1) using the combined minimum redundancy maximum relevancy criterion with the sparse support vector machine. *SAR and QSAR in Environmental Research*, 29:517–527.

Simas, A. B., Barreto-Souza, W., and Rocha, A. V. (2010). Improved estimators for a general class of beta regression models. *Computational Statistics & Data Analysis*, 54(2):348–366.

Xia, X., Gui, L., He, G., Xie, C., Wei, B., Xing, Y., Wu, R., and Tang, Y. (2017). A hybrid optimizer based on firefly algorithm and particle swarm optimization algorithm. *Journal of Computational Science*.

Zhao, W., Zhang, R., Lv, Y., and Liu, J. (2014). Variable selection for varying dispersion

beta regression model. *Journal of Applied Statistics*, 41(1):95–108.