MJFAS MALAYSIAN JOURNAL OF FUNDAMENTAL AND APPLIED SCIENCES
PRINT ISSN: 2289-5981 | ONLINE ISSN: 2289-599X

RESEARCH ARTICLE

Improvement of estimation based on small number of events per variable (EPV) using bootstrap logistics regression model

Muhamad Safiih Lola [a, d,*], Nurul Hila Zainuddin [b], Mohd Noor Afiq Ramlee [a, d], Muhamad Na'eim Abdul Rahman [a], Mohd Tajuddin Abdullah [c, d]

[a] School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia
[b] Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, 35900 Tanjong Malim, Perak, Malaysia
[c] School of Marine Science and Environment, Universiti Malaysia Terengganu, 21030 Kuala Terengganu, Terengganu, Malaysia
[d] Kenyir Research Institute, Universisti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia

* Corresponding author: safiihmd@umt.edu.my

Article history
Received 14 June 2017
Accepted 15 December 2017

Abstract

In this research, a bootstrap approach model is proposed, namely as Bootstrap Logistics Regression Model (BLRM) that is specifically used to solve the small events per variable (EPV) problem. Considering a sample data from study case of endemic dengue at several localities in Kelantan, Malaysia, a simulation study is conducted. We generated 5, 10, 20 and 25 mean samples with 500 times replacement, 1500 times bootstrap for each small EPV value (EPV= 2, 3, 4 and 5) according to the basic reproduction number, R_0 for endemic dengue. The performance of the propose BLRM revealed that the frequency distribution of estimated regression coefficient became less peaked and possessed thinner tails; the average percent relative bias consistently decreased and was closed to true parameter; the sample variance (MSE and RMSE) of the estimated regression coefficients of were smaller than original model

Keywords: Endemic dengue, Logistics regression, Bootstrap approach, Monte Carlo simulation

INTRODUCTION

For the past thirty decades, the logistic regression has been widely used and received considerable attention from researchers and practitioners. This model has been shown to be successful when used for studying the relation between response and two or more predictor regression models as well as for modeling dichotomous outcomes.

Specifically, the logistic regression model is usually formulated mathematically based on the functional form of a logistic, and cumulative density function (*cdf*) related to the probability of the occurrence of a particular event, *E*, with a conditional on a vector, *x*, of explanatory variables, to the vector *x*. In this paper, we consider the logistic regression model which has been used for estimating EPV by Peduzzi *et al*. (1996) and define it as follows:

$$P(\text{Death}|X_i)=\left\{1+e^{\left(\ln[K_1P(\text{Death})|K_2P(\text{Survival})]\right)+X_i\beta}\right\}^{-1} \qquad (1)$$

where the conditioning on the deaths, $P(\text{Death}|\ X_i)$ is given prior consideration in dengue case, with the patient of i is noted as variable X_i. The K represents the indicator of probability the death (indicator K_1) and survivor (indicator K_2). While, the β is noted as the of logistic coefficient for each X_i .

Directly related to the model in Eq. (1) is number of events per variable (EPV). This is very common among researchers as criteria in multivariable analysis (see Peduzzi *et al*. 1985; Peduzzi *et al*., 1996; Gareth *et al*., 2002; Rahim *et al*.,2007; and Wynants *et al*., 2015). An ideal minimum EPV value is suggested in the range of 10-20 (Harrel *et al*., 1985). However, the usage of this model in Eq. (1) especially when referring to EPV values might cause other serious problems and potential misleading associations such as inaccuracy and imprecision of the regression coefficient due to small EPV values compared to the free variables in the model proposed by Concato *et al*. (1993). In the research by Peduzzi *et al*. (1985), Peduzzi *et al*.(1996) and Freedman and Pee (1989), it was stated that when the EPV value does not follow the expected minimum values set, three types of errors, which are over-fitting (Type I error), under-fitting (Type II error) and paradoxical fitting (Type III error), would occur. The identification of these errors led to the introduction of a general guideline of a minimum EPV in multivariable analysis. The study concludes that a certain number of EPVs are needed so that the validity of the model can be trusted (see for example, Peduzzi *et al*. 1985; Harrel *et al*.,1985; Concato *et al*.,1993; Concato *et al*.,1995; and Peduzzi *et al*., 1996).

However, the parameters estimate of logistic regression model revealed that bias occurred in positive and negative directions (Peduzzi *et al*., 1996), especially when EPV values are fewer than 10. There are no problems when EPV values are greater than 10. In other words, small EPV values (fewer than 10) can lead to major problems which are overestimation and underestimation. Additionally, it gives a negative impact on the validity and reliability of the logistic regression model.

Thus, to overcome these critical problems and simultaneously enhance the capabilities and performance of the LRM, this study

proposes a combination of the bootstrap method with a logistic regression model, a hybrid coined as the Bootstrap Logistics Regression Model (BLRM). The bootstrap method, initiated by Efron (1979) is a method in the computer-based nonparametric family designed to set the standard accurate measurement of an estimated sample. To investigate the effectiveness of the proposed BLRM, we used the unbiased and efficient estimator characteristics as applied by Arthur (1962), Tao and Narayanaswamy (2008), Muhamad Safiih (2013), Muhamad Safiih *et al*. (2014) and Muhamad Safiih *et al*. (2016). Investigation towards the developed model focuses on creating unbiased estimator value and small error value as well as shorter interval average compared to the original model. Through this proposed model, we constructed a standard error as well as a confidence interval of the LRM. For this purpose, we conducted a Monte Carlo simulation through R programming using data from endemic dengue fever that has had 4 deaths (events) among 320 patients. For the analysis, the complete data for the variables were taken from 15 localities, from which 4 patients died in 2009, thus yielding an EPV of (11/3=) 3.67 ≈ 4.00 for the full sample. This research used 3 types of variables, producing an EPV of 4/3 = 1.333 for the full sample. The Monte Carlo simulation was conducted for small EPV values i.e., 2, 3, 4 and 5. Finally, the results from the proposed model were compared with the original model based on bias, precision and significance testing on the regression coefficients. To explainmore detailedabout the study, we first introduced theBLRM in Section 2. In Section 3, a numerical example that illustrates the BLRM and its comparison with LRM will be presented. In this section, we used the Monte Carlo simulation study on endemic dengue fever data. To measure the effectiveness of the proposed model, standard statistical performance criteria such as bias, mean square error, root mean square error, and confidence interval were also examined. This paper is finished with a conclusion in Section 4.

METHODOLOGY

The Bootstrap Logistic Regression Model

The logistic regression model or LRM is widely applied in measuring relationships between two or more predictors. Although this type of model is widely used, misclassification estimates can still happen (Carroll and Pederson, 1993) and it becomes more challenging when a small-sized sample is involved (Peduzzi *et al*., 1996). The best alternative to solve this problem is the bootstrap method, a type of nonparametric statistical inference approach which was introduced by (Efron, and Tibshirani, 1993). Bootstrapping enhances the capacity and performance of the LRM, as we will show later using our developed BLRM. By applying this method, improved standard error or confidence interval can be developed. This is because the bigger the error value, the further the estimator from the real value and vice versa; while the smaller the confidence interval, the better the estimator value and vice versa.

To look at the effectiveness of the BLRM model towards EPV that is caused by small data sample size, a Monte Carlo experiment was conducted. Using the bootstrap method, estimation for the sample distribution could be done to almost all samples. Bootstrap method is the sampling from sample replacement, which is done by taking random samples from the original sample. Bootstrap sampling depends on the sample itself, depending on the number of sources possessed. Bootstrap equality principle stated that the estimator for sub sampling (bootstrap method) is the same as the sample estimator. In addition, other than being a more accurate sample estimator, the bootstrap method can measure variability and bias. The bootstrap concept can be explained through Fig. 1(a) and Fig. 1(b) which represent real data and bootstrap data. Based on Fig. 1(a), we assume that probability distribution is unknown, *F*, gives data obtained as $x=(x_1, x_2, ..., x_n)$ through random sampling, while measurable statistic uses x, $\hat{\vartheta}=s(x)$. Fig. 1(b), which is the bootstrap data (in this study, we use Freedmen's term, which is referred to Rahim *et al*., 2007, as empirical distribution), $\hat{F}$ giving bootstrap samples $x^*=(x_1^*, x_2^*, ..., x_n^*)$ by random sampling that is measured from bootstrap statistic replication, $\hat{\vartheta}^*=s(x^*)$. The advantage of using bootstrap data is that we can calculate as many replications of $\hat{\vartheta}^*$ as possible. The calculation of $\hat{F}$ from *F* is shown by the big white colored arrow. These concepts are an important part in the bootstrap process (Efron and Tibshirani, 1993).

Algorithm of Experimental Study

The algorithm of the LRM model using the bootstrap approach as proposed is based on Fig. 1(a) and Fig. 1(b). The steps are as follows:

Step 1: Prior to the Monte Carlo simulation, death selection is calculated, i.e. *j*. Three types of variables were present in this research, i.e. x_1, x_2 and x_3 with total dengue fever cases, average temperature and total number of mosquito breeding respectively. The death selection is represented by $j=3\times \mathrm{EPV}$. Thus, EPV for the full sample can be calculated using the equation:

$$\mathrm{EPV\ for\ full\ sample} = \mathrm{Total\ Variables} \times \frac{\mathrm{Total\ Death}}{\mathrm{Total\ Variables}}$$
$$= 3 \times \frac{4}{3}$$
$$= 4.0$$

In conclusion, EPV for the full sample of dengue case study in this research is 4.0.

Step 2: Generate 500 simulations using sampling with replacement, where each simulation data uses fixed EPV value of 2, 3, 4, or 5. For example, the first 500 generated values used an EPV of 2, and the third 500 generated values used an EPV of 4. Fifteen endemic dengue fever localities were chosen for simulations.

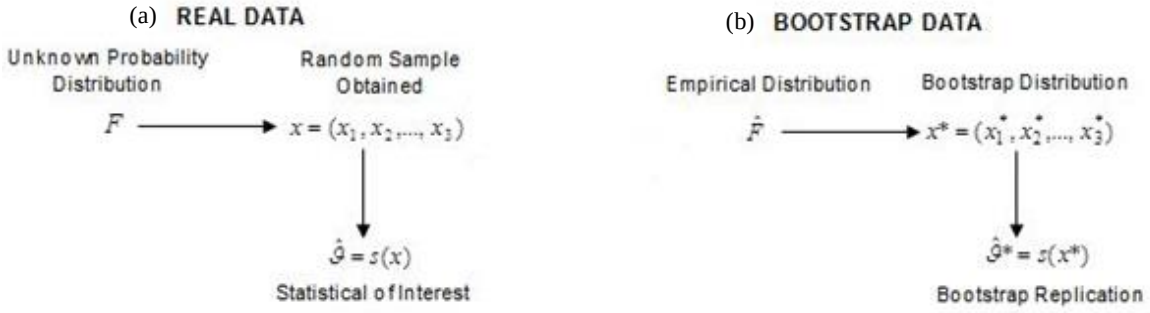

Fig. 1 (a) Real data and (b) Bootstrap data

Step 3: Estimate prediction probability of dying (P_i) using the logistic model (LR):

$$P_i = \{1+e^{[-(\alpha+X_i\beta)]}\}^{-1} \quad (2)$$

where α is intercept term; $X_i = (X_{i1}, X_{i2}, X_{i3})$ is the set for covariate values for patient i; and $\beta = (\beta_1, \beta_2, \beta_3)$ is the set of corresponding values of the regression coefficients estimated from the full sample with EPV = 4 as shown in Step 2.

Step 4: Calculate the residual model for LR based on $e_i = \text{death-}P_i$

Step 5: Selection of cumulative death and survival selection probabilities, $C_j(\text{death}) = \sum_{\{k=1,i\}} S_k(\text{death})$ and $C_j(\text{survival}) = \sum_{\{k=1,i\}} S_k(\text{survival})$ respectively.

Step 6: Generating uniform random numbers, $u = (u_1,..., u_i)$ between 0 and 1, with death selection, j. This process is repeated until the required number 3 x EPV death is obtained:

EPV 2	EPV 3	EPV 4	EPV 5
j = 3 x EPV = 3 x 2 = 6	j = 3 x EPV = 3 x 3 = 9	j = 3 x EPV = 3 x 4 = 12	j = 3 x EPV = 3 x 5 = 15

The selection of death is continued until the required number of 3 x EPV death is obtained for every 500 generated simulations.

Step 7: For every generated data, residual value that is obtained by LR model as shown in Step 4 is calculated using the bootstrap method. Due to this, a new bootstrap value, $e_i^{B(t)}$ can be obtained, where $i = 1,..., m$ refers to the ith time, and $t = 1,..., B$. The notation B is referred to total bootstrap replication sets, i.e. 1500.

$$\hat{e}_i^{B(t)} = \begin{bmatrix} \hat{e}_1^{B(1)} & \cdots & \hat{e}_1^{B(1499)} & \hat{e}_1^{B(1500)} \\ \vdots & & \vdots & \vdots \\ \hat{e}_{m-1}^{B(1)} & \cdots & \hat{e}_{m-1}^{B(1499)} & \hat{e}_{m-1}^{B(1500)} \\ \hat{e}_m^{B(1)} & \cdots & \hat{e}_m^{B(1499)} & \hat{e}_m^{B(1500)} \end{bmatrix}$$

Step 8: The estimation obtained in Step 7 is revaluated using Step 4 formula, thus, a bootstrap data set can be obtained as follows:

$$x_i^{B(t)} = \begin{bmatrix} x_1^{B(1)} & \cdots & x_1^{B(1499)} & x_1^{B(1500)} \\ \vdots & & \vdots & \vdots \\ x_{m-1}^{B(1)} & \cdots & x_{m-1}^{B(1499)} & x_{m-1}^{B(1500)} \\ x_m^{B(1)} & \cdots & x_m^{B(1499)} & x_m^{B(1500)} \end{bmatrix}$$

Step 9: A bootstrap sample, x_i^B can be obtained by averaging the each column of bootstrap data set in Step 9.

Step 10: Using the same calculation to estimate the value of parameter β using the bootstrap data in Step 9 and MLR model, the BMLR hybrid model is subsequently formulated in this step.

Step 11: Step 5 is repeated using bootstrap data. For Step 6, random uniform data was equal.

Based on these developed algorithms, the Bootstrap Logistic Regression Model (BLRM) is now defined as:

$$P\left(\text{Death}^B \middle| X^B\right) = \left\{1+e^{\left(\ln[K_1 P(\text{Death}^B)\middle| K_2 P(\text{Survival}^B)]\right)+X_i^B\beta}\right\}^{-1} \quad (3)$$

Statistical Performance

As mentioned earlier, the problems faced by Peduzzi *et al*. (1996) included overestimation and underestimation of EPV due to small sample sizes. To test this, we created a simulation series of small EPV values i.e. 2, 3, 4 and 5. In order to see how effective the propose model, a comparison of performance with original model is made. We used statistical performance criteria to measure the consistence, efficiency, accuracy,and predictability of the developed model as measured by bias, Mean Square Error (MSE) and confidence interval. Without loss of generality of estimation theory, this research considers two (2) standard statistics indicators which have been applied before by Peduzzi *et al*. (1996). Firstly, analysis and calculation methods were based on analysis and calculation methods used by Peduzzi *et al*. (1996). This includes (i) examining the distribution of the regression coefficient and measuring normality test using Kolmogorov Smirnov; (ii) assessing the accuracy of coefficients by calculating the average percent of relative bias used by Peduzzi *et al*. (1996), where regression coefficient for each of the $j = 1,2,3$ and each of the $k = 1,..., K$ simulation that converged is as follows:

$$100\left|\frac{\sum_{k=1}^{K}(\beta_{jk} - K\beta_{j,\text{true}})}{K\beta_{j,\text{true}}}\right| \quad (4)$$

where $\beta_{j,\text{true}}$ is the "true" value of the coefficient obtained from the full sample. Additional calculation to calculate accuracy is by the proportion of simulations in which the bias exceeded $\pm 100\%$, (iii) examining observed coefficient's accuracy and efficiency through sample variance (assumed to be MSE) calculation before comparing between "sample" and "proposed model" for every regression coefficient. This is done using the following equation:

$$\left|\frac{\sum_{k=1}^{K}(\beta_{jk} - \bar{\beta})^2}{K-1}\right| \quad (5)$$

with this, variance calculation is determined through the average variance from the LRM model for every coefficient on every model K that converged, as shown in the equation below:

$$\sum_{k=1}^{K} \text{var}(\beta_{jk}) \times K^{-1} \quad (6)$$

from Eq. (5), we can calculate the Root Mean Square Error (RMSE) based on the following equation:

$$\sqrt{\sum_{\{k=1,K\}}[(\beta_{jk} - \bar{\beta})^2 (K-1)^{-1}]} \quad (7)$$

Using bootstrap method, estimation for a sample distribution can be done for almost all statistical models. Bootstrap is sampling with replacements from sample, which is done by taking random samples from the real sample. Bootstrap sampling depends on the sample itself following the number of resources that are available. Bootstrap equality principle stated that the sub sampling is equal to the sample estimator. In addition, bootstrap method can measure variability and bias and also give accurate sample estimations.For the assessment of the accuracy of coefficients, we followed what was done by Peduzzi *et al*. (1996), i.e. the average percent relative bias through calculations for each of the $j = 1,2,3$ regression coefficient. Secondly, for the

method of evaluation, we used the same statistical significance of the regression coefficients. However, we considered only three out of four ways for the method of evaluation i.e. confidence intervals are determined as the proportion of simulations following the given equation:

$$\exp[\beta_j \pm z(1-\tfrac{\alpha}{2})s(\beta_j)] \quad (8)$$

where the value of $z(1-\frac{\alpha}{2}) = 1.645$, i.e. 95% CI for $\exp(\beta_j)$, and $s(\beta_j)$ refers to standard error of β_j. Proportion of simulations is defined as the coefficient divided by its standard error, and popularly describe as "*paradoxical fitting*".

Data

In this research, the simulated data concerns endemic dengue provided by the Health Department in Kelantan, Malaysia. There are 15 localities with endemic dengue with a total of 323 dengue fever patients between years 2005-2009. The complete data with 3 variables were available from the 15 localities. From this sample, 4 died in 2009, yielding an EPV of (12/3) = 4.00 for the full sample. The variables selected for the simulations are total number of dengue cases reported in each locality (x_1), average temperature (x_2) and total number of mosquito breeding, $R_0(x_3)$ whilst response variable (y) is the probability of deaths. Table 1(a) and Table 1(b) summarize the results of multivariate logistic model and multivariate bootstrap logistic model applied to the full sample.

Table 1(a) Statistical summary of baseline risk factor in original complete group.

Factor	Multivariable Logistic Regression Estimates			
	Coefficients	Standard error	Wald *p*-value	Odds ratio
Intercept	130.679	275.117	0.475	5.6E+56
Total reported case	0.161	0.140	1.151	1.175
Temperature	0.262	1.929	0.136	1.300
Reproduction basic number	-140.362	256.253	-0.508	1.1E-61

Table 1(b) Statistical summary of baseline risk factor in bootstrap complete group.

Factor	Multivariable Logistic Regression Estimates			
	Coefficients	Standard error	Wald *p*-value	Odds ratio
Intercept	70.008	802.008	0.087	2.5E+30
Total reported case	0.064	0.951	0.067	1.066
Temperature	-0.267	0.256	-1.044	0.765
Reproduction basic number	-64.176	809.667	-0.079	1.3E-28

RESULT AND DISCUSSION

Before discussing on performance of the proposed model regarding small EPV values, first step is to examine the departure from the normality of the z-statistic distribution. For this, we divided our study into three parts; (1) testing the covariance effects under $H_0 : \beta = 0$; (2) testing the skew; and (3) testing for goodness of fit using Kolmogorov Smirnov (K-S) and Shapiro-Wilk (S-W). The parts (1), (2) and (3) are shown in Table 2 and Table 3. As mentioned by [1], the departures from normality are common especially for small EPV ($EPV<10$). Thus, for LRM, the distributions of variables x_1 and x_2 are skewed to the left, while x_3 is skewed to the right. However, the BLRM was also skewed to the left for x_2 . These results are shown in Table 2.

Even though the skewed graphic of BLRM was more towards normal, it was also supported via asymptotic significance and test distribution, as the bootstrap method used was able to produce a small standard deviation. This graphical result can be viewed in Fig. 2.

The condition of normality distribution for both regression coefficients were further investigated using the Kolmogorov Smirnov (K-S) as well as the Shapiro-Wilk (S-W). According to normal measurement test, K-S and S-W are based on the assumption that the data follows a normal distribution if the values obtained for K-S are big, and for S-W, small.

From Table 3, it can be seen that the K-S values were larger than the S-W values for both LRM and BLRM. Thus, both tests proved that proposed LRM and BLRM adhered to the assumption that the data follows a normal distribution of all EPVs. However, BLRM values were bigger compared to LRM. This result is also supported by Fig. 2, where "peaked" distribution and thinner "tails" in both directions were more common at bootstrap regression coefficient as EPV decreased. Although bootstrap distribution showed one negative skewed pattern for both regression coefficient x_1 and x_2 for all EPV, the Kolmogorov test indicated that the data used was still approximate to normal distribution.

The best estimation i.e. the consistency and efficiency between BLRM and LRM especially when involving small EPV values are the primary focus in this study. Before discussing this condition in detail, we should first look into the effect of EPV on the frequency distribution for both models. This is done by comparisons of EPV frequency distribution of the values of the regression coefficient for variables from both proposed and original models. The total dengue cases recorded in each locality is depicted in Fig. 3. From Fig. 3(a), it is revealed that the smaller the EPV values (from EPV=5 to EPV=2), the more the frequency distributions of estimated regression coefficient are concentrated to normal distribution with mean = 0 (Nornadiah and Yap, 2011; Efron and Tibshirani, 1993).

In other words, as EPV decreased, the distribution became "flatter", particularly for LRM distribution, while BLRM became less peaked and has thinner tails. For example, in Table 4, which shows a single EPV number (e.g. 2), the minimum and maximum values of the bootstrap regression coefficient were -0.03 and 0.026 respectively, compared with -0.291 and -0.004 for the LRM coefficient. The standard deviation values for BLRM and LRM were 0.019 and 0.080 respectively. This indicates that frequency effect through standard error development towards proposed BLRM can enhance the achievement and efficiency besides giving more accurate results to the estimating model. In this manner, inaccurate estimation of the actual regression coefficient values is more likely to occur for the LRM method

The first condition for the best estimator is to investigate the consistency performance of our proposed model against LRM. In this case, it is the ratio to the average percent relative bias of coefficient as in Eq. (4), which is the same equation that was used by Peduzzi *et al*. (1996) and graphically illustrated in Fig. 4. The results showed that the average percent relative bias values of BLRM decreased with increased numbers of EPV. As mentioned by Peduzzi *et al*. (1996), small EPV values lead to inconsistent coefficient. However, with the used of plug-in principle of bootstrap in LRM, this problem can be solved; the average per cent relative bias of coefficient value was found to decline with increased EPV values as shown in Table 5 and Fig. 5. This decrement happened consistently and was smaller compared to LRM. For example, in BLRM with the coefficient value of β_2 (LRM values are in parentheses) the EPV value of 2 is 34.603 (73.408), and it continuously reduced to 28.306 (73.415), 22.120 (73.444) and 12.161 (73.445) when EPV values increased to 3, 4 and 5. These results provided evidence that the coefficient values of LRM are unstable. This table also revealed that at 2 EPV, the average percent relative bias is the highest, so that regression coefficients are overestimated by an average of 7.01% for β_1 and by 34.60% for β_2 .

Table 2 Statistical summary of z-value for LRM and BLRM coefficient of all factors.

Statistical Estimation	Variables	Events Per Variable (EPV)							
		LRM				BLRM			
		2	3	4	5	2	3	4	5
Standard deviation	x_1	0.512	0.512	0.511	0.512	0.070	0.102	0.063	0.090
	x_2	0.745	0.743	0.741	0.745	0.025	0.031	0.018	0.020
	x_3	0.326	0.327	0.326	0.326	0.019	0.027	0.018	0.029
Skewness	x_1	0.371	0.368	0.368	0.367	1.981	2.215	2.157	1.998
	x_2	-1.971	-1.975	-1.982	-1.970	-0.824	-0.558	-0.604	-0.334
	x_3	-0.965	-0.952	-0.960	0.966	0.341	0.281	0.218	1.286
p-value	x_1	0.997	0.897	0.952	0.977	0.113	0.107	0.020	0.015
	x_2	8.5E-7	3.6E-7	9.6E-8	3.8E-8	1.0E-7	2.2E-7	8.6E-8	0.015
	x_3	0.157	0.138	0.116	0.116	0.382	0.672	0.242	2.0E-5
Wald	x_1	0.00	0.016	0.003	0.00	2.510	2.596	5.409	5.842
	x_2	24.241	25.859	28.451	30.240	28.206	26.796	28.654	5.901
	x_3	1.997	2.190	2.460	2.463	0.762	0.178	1.368	18.177

Table 3 Normality test for LRM and BLRM.

Normality Test	Variables	Events Per Variable (EPV)							
		LRM				BLRM			
		2	3	4	5	2	3	4	5
Kolmogorov Smirnov Z	x_1	4.442	4.592	4.791	4.989	4.333	4.162	3.416	3.382
	x_2	4.745	4.919	5.169	5.344	6.388	5.657	4.593	4.814
	x_3	5.786	5.973	6.296	6.494	5.920	4.578	5.771	6.044
Kolmogorov Smirnov	x_1	0.150	0.150	0.149	0.149	0.145	0.131	0.092	0.085
	x_2	0.163	0.164	0.164	0.164	0.237	0.196	0.140	0.144
	x_3	0.210	0.209	0.209	0.209	0.216	0.148	0.189	0.191
Shapiro-Wilk	x_1	0.897	0.897	0.897	0.897	0.963	0.976	0.975	0.978
	x_2	0.954	0.954	0.954	0.953	0.941	0.961	0.973	0.960
	x_3	0.965	0.965	0.965	0.965	0.959	0.959	0.947	0.938
p-value	x_1	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
	x_2	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
	x_3	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01

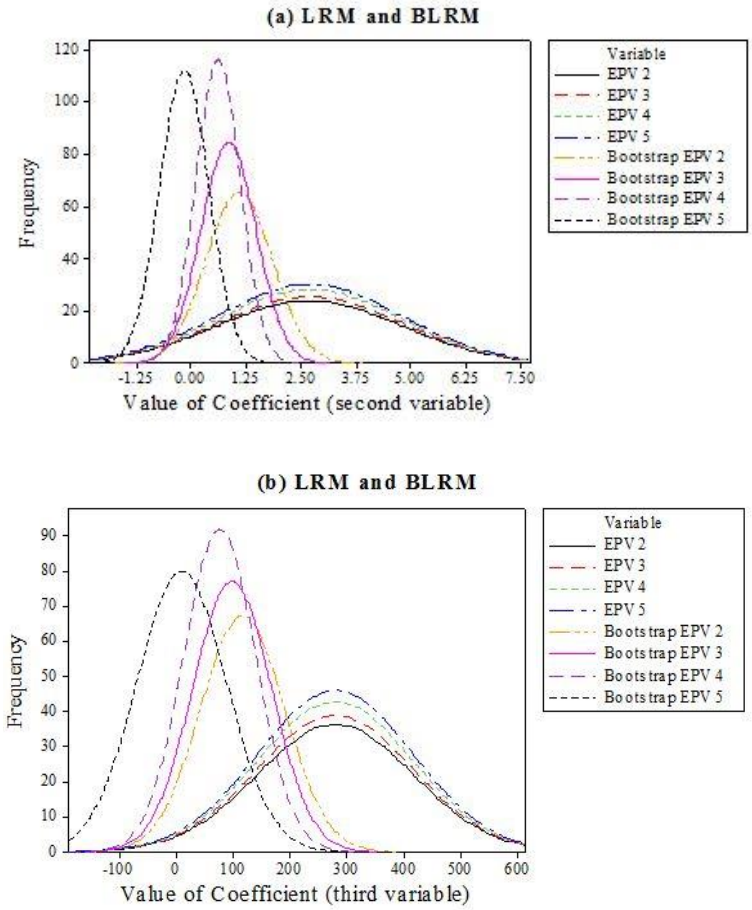

Fig. 2 Frequency distribution of regression coefficient estimation for (a) Temperature, second variable (x_2) and (b) Total basic number of reproduction, third variable (x_3).

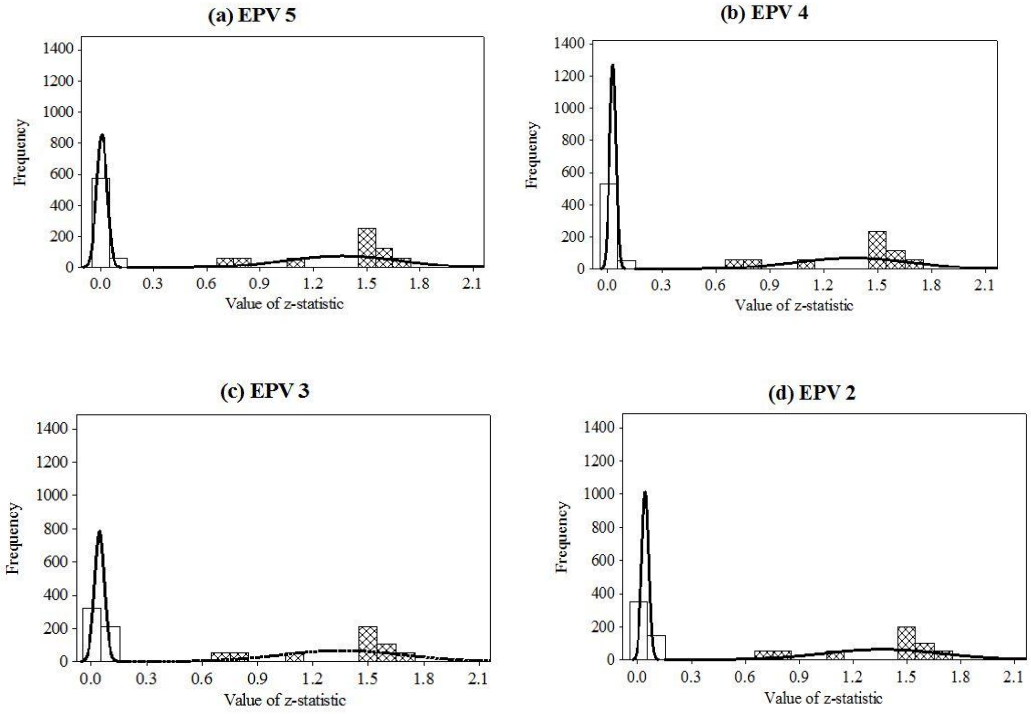

Fig. 3 Distributionof the z-statistic for reproduction number under the null hypothesis that the covariates has no effect with outcome where the box line histogram is non-hibrid distribution whereas the white histogram represents bootstrap distribution

Table 4 Statistical summary of variable x_1.

	Events Per Variable (EPV)							
	LRM				BLRM			
	2	3	4	5	2	3	4	5
Standard Deviation	0.080	0.080	0.080	0.080	0.019	0.019	0.015	0.016
Minimum	-0.291	-0.291	-0.291	-0.291	-0.030	-0.028	-0.017	-0.012
Maximum	-0.004	-0.004	-0.004	-0.004	0.026	0.032	0.032	0.038

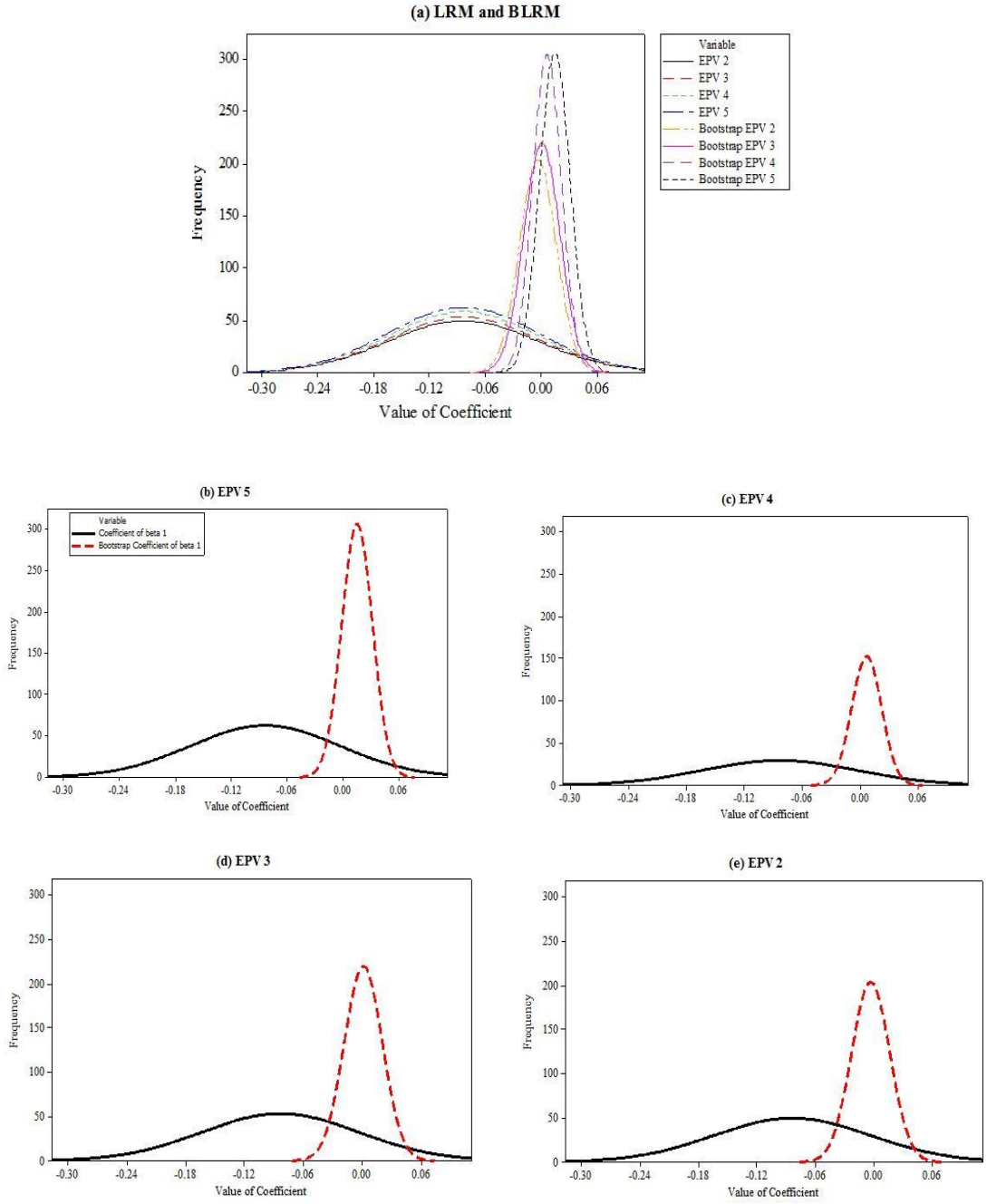

Fig. 4 Number of events per variable and frequency distribution of estimated regression coefficient for total reported dengue cases according to locality (x_1) for (a) LRM versud BLRM, (b) EPV 5, (c) EPV 4, (d) EPV 3 and (e) EPV 2

Apart from that, the consistency condition was further investigated using proportions as depicted inTable 6. At 2 EPV to EPV 5, it was revealed that the proportions decreased substantially and it exceeded 0.010 for all variables in both LRM and BLRM. However, proportions of all variables in BLRM were smaller than LRM. For instance, the proportion of simulation in which the average relative bias exceeded 100% at 2 EPV is 0.014, in contrast with 0.020 in LRM.

The odd ratio for β_1 showed that the values of BLRM were persistently greater compared to two other variables at 2 EPV, most probably due to the relatively small impact of β_2 on outcome (Table 6 and Fig. 6). For comparison, odd ratios (Table 7) for β_2 and β_3 were 3.728 and 6.1E+1, respectively. Similar patterns were observed for comparison between β_2 and β_1 .

Table 5 Average percent relative bias (100%) for LRM and BLRM.

Coefficient	Model	Events Per Variable (EPV)			
		2	3	4	5
β_1	LRM	10.126	10.117	10.111	10.121
	(BLRM)	(7.008)	(6.561)	(5.994)	(5.121)
β_2	LRM	73.408	73.415	73.444	73.445
	(BLRM)	(34.603)	(28.306)	(22.120)	(12.161)
β_3	LRM	19.937	19.930	19.924	19.951
	(BLRM)	(18.748)	(16.837)	(14.585)	(8.889)

Table 6 Proportion of simulation in which average relatives bias exceeded 100%.

Coefficient	Model	Events Per Variable (EPV)			5
		2	3		0.016
β_1	LRM	0.020	0.018	0.017	(0.008)
	(BLRM)	(0.014)	(0.012)	(0.010)	0.116
β_2	LRM	0.146	0.137	0.124	(0.019)
	(BLRM)	(0.069)	(0.052)	(0.037)	0.031
β_3	LRM	0.039	0.037	0.033	(0.014)
	(BLRM)	(0.037)	(0.031)	(0.024)	

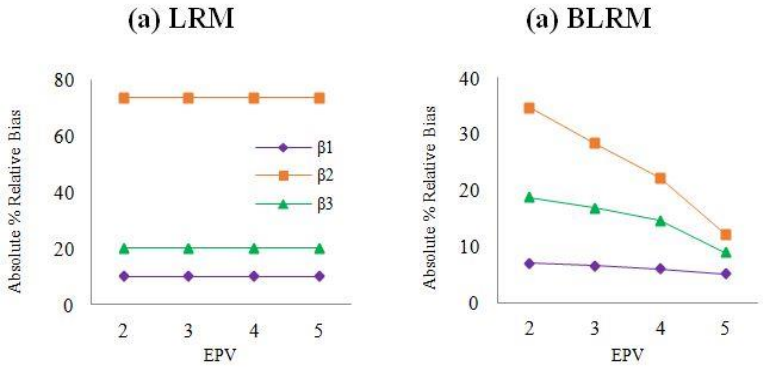

Fig. 5 Number of events per variable and average percent of relative bias for (a) LRM and (b) BLRM. Abbreviations for variables are β_1 = Totalreported dengue cases, β_2 = Locality temperature, β_3 = Aedes mosquito's reproduction estimation in the locality observed.

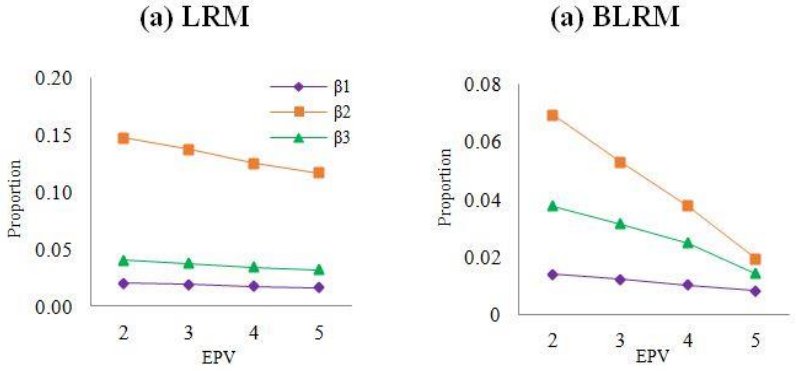

Fig. 6 Number of events per variable and proportion of simulation in which the average percent relative bias exceeded 100% , (a) LRM and (b) BLRM. Abbreviation are as indicated in Fig. 5.

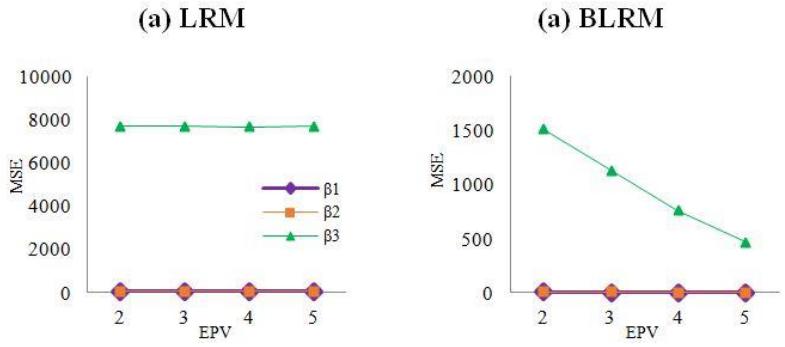

Fig. 7 Number of events per variable and MSE of (a) LRM and (b) BLRM.

Table 7 Odd ratio of LRM and BLRM.

Coefficient	Model	Events Per Variable (EPV)			
		2	3	4	5
β_1	LRM	0.922	0.922	0.922	0.922
	(BLRM)	(0.966)	(0.001)	(1.006)	(1.015)
β_2	LRM	48.482	48.630	48.616	48.498
	(BLRM)	(3.728)	(2.766)	(2.052)	(0.967)
β_3	LRM	3.3E+2	3.2E+9	3.2E+2	3.3E+2
	(BLRM)	(6.1E+1)	(9.1E+9)	(8.9E+8)	(4.7E8)

The performance of the proposed model was further investigated by looking at the efficiency conditions towards small EPV values compared to the LRM model. Our analysis concentrated on MSE and RMSE as in Eq. (6) and Eq. (7) respectively. The MSE and RMSE of regression coefficient values are displayed in Table 8 and Fig. 7. According to the estimation theory, the smaller the variance (as well as MSE and RMSE), the better the estimate would be. As illustrated in Table 8, it is clear that the MSE values of BLRM (RMSE in parentheses) decreased as the EPV values increased.

However, when we look at LRM, it is revealed that the values of MSE and RMSE were unstable as EPV values increased. For example, the MSE and RMSE (in parentheses) values of parameter β_1 for BRLM were 7.208 (2.357), 4.756 (1.891), 2.750 (1.422) and 0.959 (0.801), which decreased as EPV values increased, Compared to LRM, whose values were unstable, i.e. 40.879 (5.840), 40.926 (5.842), 40.948 (5.845) and 40.912 (5.843) for EPV values of 2, 3, 4 and 5 respectively; the BLRM values were stable. From Table 6, the results clearly showed that our proposed model was more efficient when compared to LRM, particularly when the EPV values were small. The value of MSE and RMSE (Fig. 7) were subsequently transformed into diagrams. It was found that all coefficients showed a down ward pattern.

To prove that the proposed model is better in estimating when extended further through confidence interval, Eq. (8) is used. According to the estimation theory, the shorter the interval, the better the estimate or by definition of confidence interval, i.e. the level of % of the true values included in the model.Table 9 and Fig. 8 show that the proportion of simulations in which 95% confidence limit about estimated value included the true value for both models. Under coverage occurred with the greatest variability in coverage for LRM estimation for all variables. Table 9 shows that all coefficient values of BRLM were greater than 95%. On the other hand, the coefficient values for LRM were as follows: β_1 a bit more than 40%; β_2 and β_3 are 35% and less than 25%, respectively. This result revealed that our proposed model included 90%-95% of the true values which were included in the model.However, based on Table 9,only 25 % to 40% of the true values for LRM were included in the model (or less than half of BRLM). The proportion of bootstrap of β_2 at 2 EPV and β_3 at 3 EPV were still greater than 0.90, and can be considered to be in the range of reliable and high coverage of confidence interval (Peduzzi *et al*., 1996). From this result, it is clear that by hybridizing the bootstrap method with LRM, the model effectiveness can be increased especially when EPV values are small.

Table 8 Comparison between LRM and BLRM based in MSE and RMSE.

Coefficient	Model	Events Per Variable (EPV)							
		MSE				RMSE			
		2	3	4	5	2	3	4	5
β_1	LRM	40.87	40.92	40.94	40.91	5.84	5.84	5.84	5.84
	(BLRM)	(7.20)	(4.75)	(2.75)	(0.95)	(2.35)	(1.89)	(1.42)	(0.80)
β_2	LRM	53.41	53.48	53.50	53.46	6.67	6.67	6.68	6.67
	(BLRM)	(9.22)	(6.04)	(3.47)	(1.29)	(2.67)	(2.13)	(1.59)	(0.92)
β_3	LRM	7688.71	7686.84	7678.39	7703.07	79.82	79.78	79.78	79.89
	(BLRM)	(1510.07)	(1129.47)	(760.34)	(464.88)	(33.32)	(27.96)	(27.96)	(14.26)

Table 9 Proportion of simulation with 95% confidence interval of the estimated regression coefficient.

Coefficient	Model	Events Per Variable (EPV)			
		2	3	4	5
β_1	LRM	0.402	0.401	0.402	0.401
	(BLRM)	(0.963)	(0.949)	(0.964)	(0.944)
β_2	LRM	0.350	0.350	0.350	0.351
	(BLRM)	(0.906)	(0.905)	(0.915)	(0.917)
β_3	LRM	0.246	0.246	0.246	0.246
	(BLRM)	(0.964)	(0.962)	(0.974)	(0.980)

Table 10 Proportion of simulation in which the square root of Wald statistics exceeded the standard normal deviate of 1.645.

Coefficient	Model	Events Per Variable (EPV)			
		2	3	4	5
β_1	LRM	0.840	0.900	0.880	0.900
	(BLRM)	(0. 600)	(0.600)	(0.590)	(0.600)
β_2	LRM	0.841	0.858	0.880	0.897
	(BLRM)	(0.580)	(0.626)	(0.638)	(0.648)
β_3	LRM	0.841	0.857	0.880	0.896
	(BLRM)	(0.618)	(0.666)	(0.681)	(0.694)

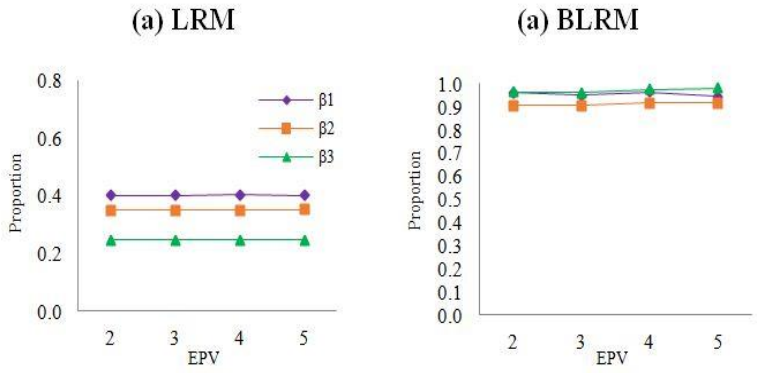

Fig. 8 Proportion of simulation in which the 95% confidence interval of the estimated regression coefficient included the true value of (a) LRM and (n) BLRM. Abbreviations are as indicated in Fig.5.

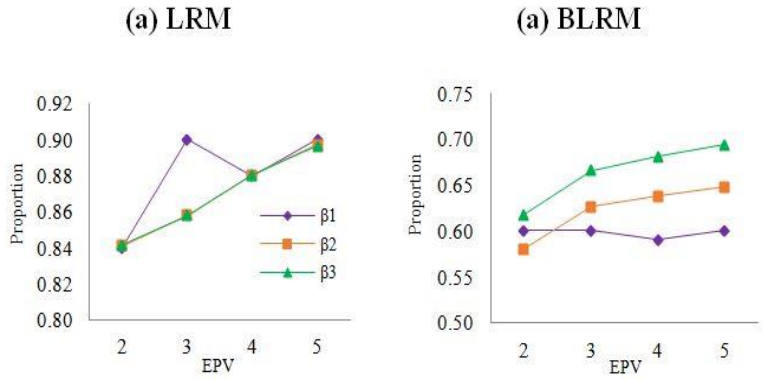

Fig. 9 Proportion of simulation in which the z-statistic (coefficient/standard error) exceeded the standard normal deviate 1.645 for 95% power of (a) LRM and (b) BLRM.Abbreviations are as indicated in Fig.3.

Table 11 Proportion of simulation in which the z-statistic < 95% level: Type III error (paradoxical fitting).

Coefficient	Model	Events Per Variable (EPV)			
		2	3	4	5
β_1	LRM	0.50080	0.50075	0.50068	0.50063
	(BLRM)	(0.42152)	(0.42147)	(0.42114)	(0.42086)
β_2	LRM	0.18459	0.18431	0.18415	0.18444
	(BLRM)	(0.42152)	(0.38316)	(0.38300)	(0.38227)
β_3	LRM	0.15914	0.15911	0.15907	0.15904
	(BLRM)	(0.38285)	(0.34527)	(0.34533)	(0.34481)

Table 10 and Fig. 9 show the proportion of simulations in which the z-statistic (square root of Wald statistic) exceeded the standard normal deviate of 1.645 which returned a 5% significance level. The power for variables bootstrap β_1 decreased slowly with increasing EPV number. However, starting from 3 EPV until 5 EPV, the proportion of bootstrap β_2 and bootstrap β_3 slowly decreased, most probably due to the greater standard errors (data not shown). For example, at 3 EPV, the BLRM for β_2 and BLRM for β_3 were 33.72 and 197.29 respectively, while the BRLM for β_1 was 0.85.

Table 11 show the proportion of simulations in which the z-statistic was paradoxically reversed to value less than -1.645, which was low for LRM, while for BLRM it was moderate for all regression coefficients except for BRLM β_1. Additionally, LRM indicated inconsistent decreasing and increasing pattern of paradoxical fitting as the EPV values rose (EPV 2 to EPV 5). For example, the pattern of β_2 started with 0.18459 then decreased to 0.18431 and 0.18415 but later increased to 0.18444. However, if we look at BRLM, the coefficient β_2 showed a decreasing pattern starting with 0.42152 before dropping to 0.38316 and declining slowly to 0.38300 and 0.38227 as the EPV increased.

CONCLUSION

This research was derived from the overestimation and underestimation issues faced by Peduzzi *et al*. (1996), especially when involving the small EPV values. To overcome these issues, we constructed a hybrid LRM with bootstrap method, namely BLRM, to improved standard error developing method and confidence interval. To measure the effectiveness of the developed model and to ensure no loss of generality, all measurements used by Peduzzi *et al*. (1996) were also used in this research. The simulation studies revealed that consistency and efficiency of the proposed BLRM could solve the problems that are insofar still faced by Peduzzi *et al*. (1996). Among them: (i) the proposed BLRM model confirmed that allproduced regression coefficient values were consistent. This can be shown by the regression coefficient value that decreased simultaneously with the increased EPV value. In contrast, regression coefficient values for LRM were unstable (labile) as increment and decrement of regression coefficient values occurred when the EPV value increased; (ii) the proposed BLRM model also revealed that regression coefficient value produced was efficient. This can be proven by MSE and RMSE values that decreased with the increment of EPV value. But this scenario did not occur for LRM where MSE and RMSE values increased simultaneously with EPV value increment.

The same result was obtained for the confidence interval, where BLRM values were shorter compared to LRM values. This indicates that the estimated value produced by BLRM is better compared to LRM especially those involving small EPV values. Overall, the hybridized bootstrap and LRM method showed a more consistent, efficient coefficient regression and produces shorter CI compared to when LRM was used as the sole method. According to this research, it is clear that the suggested BLRM could overcome the problems faced by Peduzzi *et al*. (1996).

ACKNOWLEDGEMENT

The authors wish to express their utmost gratitude to the Health Department in Kelantan, Malaysia for the collaboration in providing the Dengue sample data. A special gratitude for School of Informatics and Applied Mathematic (SIAM), Kenyir Research Institute (KRI) and Research Management Centre (RMC), Universiti Malaysia Terengganu for supported this research study. We also thank Ms. Ernisa Marzuki from Universiti Malaysia Sarawak (UNIMAS) for the support and assistance during the completion of this research..

REFERENCES

Arthur, S. G. 1962. Best Linear Unbiased Prediction in the Generalized Linear Regression Model. *Journal of the American Statistical Association*, 57(298), 369-375.

Asghar, G. and Saleh, Z. 2012. Normality Tests for Statistical Analysis: A Guide for Non-Statisticians. *International Journal of Endocrinology and Metabolism*, 10(2), 486-489.

Carroll, R. J. and Pederson, S. 1993. On Robustness in the Logistic Regression Model. *Journal of the Royal Statistical Society, Series B*, ***80***, 461-465.

Concato, J., Feinstein, A. R. and Holford, T. R. 1993. The Risk of Determining Risk with Multivariable Models. *Annals of Internal Medicine*, 118, 201-210.

Concato, J., Peduzzi, P., Holford, T. R., and Feinstein, A. R. 1995. The Importance of Events per Independent Variable (EPV) in Proportional Hazards Analysis: I. Background, Goals and General Strategy. *Journal of Clinical Epidemiology*, 48, 1495-1501.

Efron, B. 1979. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1), 1-26.

Efron, B. and Tibshirani, R. J. 1993. An Introduction to the Bootstrap. New York: Chapman and Hall.

Freedman, L. S. and Pee, D. 1989. Return to a Note on Screening Regression Equations. *American Statistician*, 43, 279-282.

Gareth, A., Anthony, R. B. and Patrick, R. 2002. Simplifying a Prognostic Model: A Simulation Study Based on Clinical Data. *Statistics Medical*, 21, 3803–3822.

Harrel, F., Lee, K. L., Matchar, D. B. and Reichert, T. A. 1985. Regression Models for Prognostic Prediction: Advantages, Problems and Suggested Solutions. *Cancer Treatment Reports*, 69, 1071-1077.

Muhamad Safiih, L. 2013. Fuzzy Parametric Sample Selection Model: Monte Carlo Simulation Approach. *Journal of Statistical Computation and Simulation*, 83(6), 992-1006.

Muhamad Safiih, L., Kamil, A.A. and Abu Osman, M.T. 2014. Estimated and Analysis of the Relationship Between the Endogenous and Exogenous Variables using Fuzzy Semi-Parametric Sample Selection Model. *American Journal of Applied Sciences*, 11(9), 1542-1552

Muhamad Safiih, L., Wan Saliha, W. A. and Nurul Hila, Z. 2016. Sample Selection Model with Bootstrap (BPSSM) Approach: Case Study of the Malaysian Population and Family Survey. *Open Journal of Statistics*, **6**, 741 – 748.

Nornadiah, M.R. and Yap, B.W. 2011. Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests. *Journal of Statistical Modeling and Analytics*, 2(1), 21-33.

Peduzzi, P., Concato, J., Kemper, E., Theodore, R.H. and Alvan, R.F. 1996. Simulation Study of the Number of Events per Variable in Logistic

Regression Analysis, *Journal of Clinical Epidemiology*, 49(12), 1373-1379.

Peduzzi, P., Detre, K. and Gage, A. 1985. Veterans Administration Cooperative Study of Medical Versus Surgical Treatment for Stable Angina-Progress Report: Section 2-Design and Baseline Characteristics. *Progress in Cardiovascular Diseases*, 28, 235-243. 1974.

Rahim, M., Flora, I.M. and Richard, H.G. 2007. A Simulation Study of Sample Size for Multilevel Logistic Regression Models. *Medical Research Methodology*,7, 34.

Tao, L. and Narayanaswamy, B. 2008. Best Linear Unbiased Estimators of Parameters of a Simple Linear Regression Model Based on Ordered Ranked Set Samples. *Journal of Statistical Computation and Simulation*, 78(12), 1267-1278.

Wynants, L., Bouwmeester, W., Moons, K.G.M., Moerbeek, M., Timmerman, D. and Van, S., Vergouwe, Y. 2015. A simulation Study of Sampel Siza Demonstrated the Importance of the Number of Events Per Variable to Develop Prediction Models in Clustered Data. *Journal of Clinical Epidemiology*, 68, 1406-1414.