Websites:
http://www.lifesciencesite.com
http://www.sciencepub.net

Emails:
editor@sciencepub.net
sciencepub@gmail.com

Life Science Journal

MARSLAND PRESS
Multidisciplinary Academic Journal Publisher

Robust Bootstrap Procedure for Estimation of Binary Logistic Regression Model in the Presence of High Leverage Points with Medical Applications

Syaiba Balqish Ariffin [1], Habshah Midi [1], A.H.M. Rahmatullah Imon [2]

[1.] Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia
[2.] Department of Mathematical Sciences, Ball State University, Muncie, Indiana, USA
syaibabalqish@gmail.com

Abstract: The classical bootstrap method should be used with caution in binary logistic regression model since it can be easily affected by high leverage points. As a remedy to this problem, we propose two robust bootstrap methods, called the diagnostic logistic before bootstrap (DLGBB) and the weighted logistic bootstrap with probability (WLGBP). In the DLGBB, the high leverage points are excluded before applying the resampling process, and for the WLGBP, the high leverage points are attributed with low probabilities to be selected in the resampling process. The usefulness of our proposed methods is investigated through medical data and simulation study. Both the empirical and simulation results confirm that the DLGBB and the WLGBP methods give significant improvement over the classical bootstrap method.

[Ariffin SB, Midi H, Imon AHMR. **Robust Bootstrap Procedure for Estimation of Binary Logistic Regression Model in the Presence of High Leverage Points with Medical Applications**. *Life Sci J* 2021;18(9):45-62]ISSN1097-8135 (print); ISSN2372-613X(online)http://www.lifesciencesite.com. 7.doi:10.7537/marslsj180921.07.

Keywords: Logistic regression; maximum likelihood; random *X* resampling; robust bootstrap; high leverage points

1. Introduction

Traditionally, in medical research, the prediction on probability of survival rate has been investigated by clinical trials about what medication or resource a patient might use for a given treatment. Confidence intervals have been used for many years in the reporting of clinical data to reflect the stochastic nature of data collected from a sample of patients. Several authors have explored methods for the approximation of confidence intervals, and the use of a statistical methodology known as bootstrapping has been put forward as a promising solution.

Bootstrap is a nonparametric method of constructing confidence intervals and estimating parameters and standard errors. The most attractive feature of bootstrap method is that it can produce the standard errors of any complex estimator without going through its mathematical theory. The bootstrap procedure which was proposed by Efron (1979) is a computer intensive method where a set of observations can be assumed as a population. This can be implemented by constructing a huge number of samples, each of which is obtained by random simple sampling with replacement from the original dataset (Efron, 1979; Efron and Tibshirani, 1998). Even though the bootstrap method is designed and proven to provide more satisfactory result when classical set up fails without much affecting the situation where classical method works, but caution must be taken while considering this method because there is no guarantee on the accuracy of the parameters estimation in the presence of outliers or high leverage points. It is now evident that the classical bootstrap method is easily affected by outliers. The reason is mainly because of the fact that the bootstrap samples may have more outliers than the original sample because of employing bootstrap resampling with replacement (Imon and Ali,2005; Norazan et al., 2009)

Bootstrap method with robust estimator may be employed as a remedy to this problem, but this may not be efficient when the percentage of outliers is higher than the breakdown point of the estimator. Moreover, the model structure in the bootstrap samples may change in the presence of outliers or high leverage points, and any method employed does not automatically provide valid inferential statements. Bootstrap techniques are frequently used in linear regression models. In recent years it is being applied in logistic regression as well. Likewise the linear regression model, the role of outliers is difficult to understand in logistic regression. Outliers in the *X*-space are called high leverage points which usually exert too much influence when the classical maximum likelihood estimator are employed (Syaiba and Habshah, 2010).

In this paper, we would like to investigate the performance of the classical bootstrap (CB) method in logistic regression in the presence of high leverage points. A large body of literature is now available for robust bootstrap methods in linear regression (Shao, 1992; Stromberg, 1997; Efron and Tibshirani, 1998; Salibian-Barrera and Zamar, 2002; Amado and Pires, 2004; Willems and Aelst, 2005; Imon and Ali, 2005; Salibian-Barrera, 2006; Norazan et al., 2009; Habshah et al., 2009), but to the best of our knowledge, there is not much study on the robust bootstrap methods in logistic regression. However, there are few papers that deal with bootstrapping in the logistic regression model based on the classical maximum likelihood estimator (MLE) (Izrael et al., 2002; Roberts et al., 2003; Hossain and Khan, 2004). The work of Imon and Ali (2005) and Norazan et al. (2009) have motivated us to develop two robust bootstrap methods in logistic regression that we call the diagnostic logistic before bootstrap (DLGBB) and the weighted logistic bootstrap with probability (WLGBP). In the DLGBB method, the suspected high leverage points are identified by the robust logistic diagnostic (RLGD) method (Syaiba and Habshah, 2010) and omitted from the original data before performing bootstrap analysis with the remaining good data (Imon and Ali, 2005). On the other hand, in the WLGBP method the RLGD method is applied first to compute probabilities which act as control mechanism in random resampling process, so that the high leverage points receive lower probabilities of being selected in the bootstrap resamples (Norazan et al., 2009).

2. Material and Methods

2.1 Robust Logistic Diagnostic

The identification of high leverage points in logistic regression model was first highlighted in Imon (2006). The RLGD method is then proposed to remedy the problem of swamping and masking effect (Syaiba and Habshah, 2010). Consider a general regression model with $k = p + 1$ coefficients.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon \qquad (1)$$

Suppose we have a sample of n observations. Then, Eq.(1) can be simplified in matrix form as:

$$Y = X\beta + \varepsilon \qquad (2)$$

where Y is an $n \times 1$ vector of responses. In logistic regression, we would logically let $y_i = 0$ if the i^th^ unit does not have the characteristic and $y_i = 1$ if the i^th^ unit does possess that characteristics X is an $n \times k$ matrix of covariates, $\beta^T = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$ is the vector of regression coefficients and ε is an $n \times 1$ vector of unobserved random errors. The specific form of the logistic regression model we use in this paper is

$$\pi(X) = \frac{\exp(X\beta)}{1 + \exp(X\beta)} \qquad (3)$$

Here, the quantity π_i is known as probability or fitted value for the i^{th} covariate. The model given in Eq.(3) satisfies $0 \le \pi_i \le 1$. The model in terms of Y would be written as:

$$Y = \pi(X) + \varepsilon \qquad (4)$$

The fitted values in logistic regression model are calculated for each covariate which depend on the estimated probability for that covariates, denoted as $\hat{y}_i = \hat{\pi}_i$. Thus, the i^{th} residuals are defined as:

$$\hat{\varepsilon}_i = y_i - \hat{\pi}_i \qquad (5)$$

The RLGD method is utilized as an initial stage in our proposed robust bootstrap methods to identify the high leverage points. On the second stage of the RLGD method, the potential values are computed by the distance from the mean values. Denote $\hat{\beta}^{(-D)}$ as the corresponding vector of estimated coefficients after D suspected high leverage points are deleted which yields the fitted values

$$\hat{\pi}_i^{(-D)} = \frac{\exp\left(X\hat{\beta}^{(-D)}\right)}{1 + \exp\left(X\hat{\beta}^{(-D)}\right)} \qquad (6)$$

Let $\tilde{X} = V^{\frac{1}{2}} X$ where V is diagonal matrix with elements $v_i = \hat{\pi}_i(1 - \pi_i)$ .Thus, group deleted distance from mean based on group deleted cases D is

$$b_i^{(-D)} = x_i^T \left(\tilde{X}_R^T \tilde{X}_R\right)^{-1} x_i \qquad (7)$$

The value of Eq.(7) computed using $\hat{\beta}^{(-D)}$ for R remaining cases and $x_i^T = \left[1, x_{1i}, x_{2i}, \ldots, x_{pi}\right]$ is the $1 \times k$ vector of observations corresponding to the i^{th} case. The relationship between potential values proposed by Hadi (1992) and Eq.(7) gives

$$b_i^{(-D+i)} = x_i^T \left(\tilde{X}_R^T \tilde{X}_R + x_i x_i^T\right)^{-1} x_i = \frac{b_i^{(-D)}}{1 + b_i^{(-D)}} \qquad (8)$$

Based on group deleted cases indexed by D, by adopting distance from mean, we define the group deleted potential denoted by

$$p_{ii}^{*(-D)} = \begin{cases} \dfrac{b_i^{(-D)}}{1 + b_i^{(-D)}} & ; i \in R \\ b_i^{(-D)} & ; i \in D \end{cases} \qquad (9)$$

Since the distribution of $p_{ii}^{*(-D)}$ is unknown, we apply cut-off point based on median and median absolute deviation (MAD) for $p_{ii}^{*(-D)}$ as suggested by Hadi (1992). Hence, any observation corresponding to excessively large potential values with cut-off points

$$p_{ii}^{*(-D)} > Med\left(p_{ii}^{*(-D)}\right) + 3MAD\left(p_{ii}^{*(-D)}\right) \qquad (10)$$

where

$$MAD\left(p_{ii}^{*(-D)}\right) = \frac{Med\left\{\left|p_{ii}^{*(-D)} - Med\left(p_{ii}^{*(-D)}\right)\right|\right\}}{0.6745} \quad (11)$$

will be declared as high leverage points. The step of RLGD method is summarized as follows:

1. For each point, compute the robust Mahalanobis distance (RMD) defined as:
$$RMD_i = \sqrt{\left(x_i - T(X)\right)^T C(X)^{-1}\left[x_i - T(X)\right]}$$
where the estimation subset is determined by either the minimum covariance determinant (MCD) or the minimum volume ellipsoid (MVE) methods proposed by Rousseeuw and Leroy (1987). These two robust estimators (MCD and MVE) can be easily computed by standard available routines of the Robust Package of SPLUS or R.

2. The i[th] point with
$$RMD_i > Med(RMD_i) + 3MAD(RMD_i)$$
are suspected as high leverage points and will be included in the deletion set D.

3. Based on the remaining observations from the set R, compute $p_{ii}^{*(-D)}$ using Eq.(10).
4. Any deleted points with
$$p_{ii}^{*(-D)} > Median\left(p_{ii}^{*(-D)}\right) + 3MAD\left(p_{ii}^{*(-D)}\right)$$
are finally declared as high leverage points.

2.2 Classical Bootstrap

There are two different ways of generating bootstrap samples in regression namely the fixed-*X* resampling and the random-*X* resampling. The fixed-*X* resampling is also known as bootstrapping residuals. A set of residuals by the MLE method is estimated, and then the bootstrap residuals are replicated by random sampling from the estimated residuals. Unfortunately, this procedure is not suitable for logistic regression model. We have already mentioned that the most extreme points in the covariate space may have the fitted probabilities which are closer to 1 or 0, consequently may affect the estimation and produce larger bias (Croux et al., 2002; Croux and Haesbroeck, 2003). Therefore the fixed-*X* bootstrapping may be very much ineffective for logistic regression model. Another approach of generating bootstrap samples is the random-*X* resampling which is also known as the case resampling or bootstrapping pairs, where the regression model is fitted with response variable Y and explanatory variables X .The resampling procedure therefore involves sampling pairs with replacement from the original data set (Imon and Ali, 2005; Norazan et al., 2009). The following summarizes the random-*X* resampling procedure:

1. The bootstrap sample
$$(y_1, x_1)^*, (y_2, x_2)^*, \ldots, (y_n, x_n)^*$$
is taken independently with equal probabilities $1/n$ from the original sample
$$(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n) .$$
2. Compute β^* from the bootstrap sample
$$(y_1, x_1)^*, (y_2, x_2)^*, \ldots, (y_n, x_n)^* .$$
3. Repeat Step 1 and Step 2 for B times to obtain $\beta^{*1}, \beta^{*2}, \ldots, \beta^{*B}$ bootstrap replications

2.3 Robust Bootstrap

The major drawback of the classical bootstrap method is that it shows a tendency of possessing higher number of high leverage points as compared to the original sample because of employing resampling with replacement (Norazan et al., 2009; Habshah et al., 2009). Even the resampling without replacement would not be an option to guarantee that bootstrapping is free from high leverage points. This problem becomes worse when the bootstrap samples are fitted using the MLE because the MLE is easily affected even by a single high leverage point (Croux et al., 2002; Croux and Haesbroeck, 2003).

2.3.1 Diagnostic Logistic Before Bootstrap

Imon and Ali (2005) pointed out it may be useless to apply diagnostic tools after the bootstrapping is done as the bootstrap replicates may include more unusual cases than the original sample. So if any protective measure is taken it should be done before bootstrap but not after that. They proposed the diagnostic before bootstrap (DBB) in the context of linear regression model in order to protect the bootstrap method against the outliers. The suspected outliers were first identified by using the robust reweighted least squares (RLS) and then deleted from the analysis before performing bootstrap with the remaining observations. Their works have inspired us to develop a robust bootstrap method for logistic regression model in the presence of high leverage points. In the proposed method, the suspected high leverage points were first identified by the RLGD method and then omitted from the analysis. The bootstrap procedure is then applied only to good observations. Norazan et al. (2009) pointed out that the weakness of DBB is its crude rejection method by assigning weight '0' to identified outliers and weight '1' to the remaining points. By this crude rejection, there is a tendency for moderate outliers to receive '0' weight, thus leaving only a few observations for the remaining set. To overcome this shortcoming, we need to rely on a good detection method. The RLGD method is proven to be a good detection method that can identify the high leverage points correctly and free from swamping and masking effects (Syaiba and Habshah, 2010). Thus, the DLGBB method based on the RLGD method can improve the bootstrapping procedure in logistic regression model. The following summarizes the bootstrapping process for the DLGBB method:

1. Apply the RLGD using either the MCD or the MVE to n original sample
$$(y_1,x_1),(y_2,x_2),\ldots,(y_n,x_n).$$
The high leverage points thus identified will form the deletion set D
2. The bootstrap sample
$$(y_1,x_1)^*,(y_2,x_2)^*,\ldots,(y_{n-d},x_{n-d})^*$$
is taken independently with equal probabilities $1/n$ from the remaining sample
$$(y_1,x_1),(y_2,x_2),\ldots,(y_{n-d},x_{n-d}).$$
3. Compute β^* from the bootstrap sample
$$(y_1,x_1)^*,(y_2,x_2)^*,\ldots,(y_{n-d},x_{n-d})^*.$$
4. Repeat Step 2 and Step 3 for B times to obtain $\hat{\beta}^{*1},\beta^{*2},\ldots,\hat{\beta}^{*B}$ bootstrap replications

2.3.2 Weighted Logistic Bootstrap with Probability

Here we propose another bootstrap method that we call the weighted logistic bootstrap with probability (WLGBP). We extend the idea of Norazan et al. (2009) by incorporating a re-descending function, namely the Bisquare weighting function for determining weights of the observations in the original sample. Here the outlying observations are attributed with low probabilities and consequently have low chances of being selected in the resampling process. The probability of a particular point in the original sample to be selected in the resampling will be based on the assigned weights. At first we assign weights to the observations in the original sample. There are several weighting functions such as Hampel, Tanh, Huber and Tukey that have been used in robust bootstrap for linear regression model (Norazan, 2008). In logistic regression model, we specifically focus on the Bisquare weighting function in order to give less weight to the high leverage points. In the WLGBP, we first apply the RLGD method to identify the high leverage points. Based on this method, we obtain the set of weight for the weighted maximum likelihood estimator (WMLE). The weights are denoted as w_i^{RLGD} and are defined as:

$$w_i^{RLGD}=\min\left\{1,\frac{p}{\left(p_{ii}^{*(-D)}\right)^2}\right\} \qquad (12)$$

where p is the number of coefficients without intercept terms, yielding

$$r_i=y_i-\left(w_i^{RLGD}\hat{\pi}_i\right) \qquad (13)$$

The Eq.(13) was first introduced by Hubert and Rousseeuw (1997). They computed positive weights, w_i based on the robust Mahalanobis distance $RMD(x_i)$ defined as:

$$w_i=\min\left\{1,\frac{p}{\left(RMD(x_i)\right)^2}\right\} \qquad (14)$$

Thus, iterative estimates of $\hat{\beta}^{WMLE}$ are then obtained as:

$$\hat{\beta}^{(K+1)}=\beta^{(K)}+\left(X^TVX\right)^{-1}X^T\left(y_i-\hat{\pi}_i\right) \qquad (15)$$

The $\hat{\beta}^{(-D)}$ from the RLGD method is used as an initial estimated coefficients to compute V as diagonal matrix with elements $v_i=w_i^{RLGD}\hat{\pi}_i(1-\pi_i)$. The weights w_i^{RLGD} are assigned to down weight the high leverage points. We obtain the modified standardized Pearson residual (MSPR) based on the WMLE defined as:

$$r_{si}^{WMLE}=\frac{\left(y_i-\hat{\pi}_i\right)}{\sqrt{v_i\left(1-h_i\right)}} \qquad (16)$$

For the notational ease, let us define $u=r_{si}^{WMLE}$ in Eq.(16). Then we define the new weights as:

$$w(u)=\psi_{Bisquare}(u)/u \qquad (17)$$

where $\psi_{Bisquare}(u)$ is defined as:

$$\psi_{Bisquare}(u)=\begin{cases}u\left[1-(u/d)^2\right]^2 & ;abs(u)\le d\\ 0 & ;d\le abs(u)\end{cases} \qquad (18)$$

and $d=4.685$ is the tuning constant. Based on these weights, we expect that the high leverage points in the original sample will receive relatively less weights. We expect that only the excessively high leverage points will receive weight '0'. To protect the whole procedure against the high leverage points, we propose to do bootstrap resampling with selection probabilities. Thus, the ith observation will get the selection probability of p_i where

$$p_i=w_i\Big/\sum_{i=1}^{n}w_i \text{ and } 0\le p_i\le 1.$$

These selection probabilities become a control mechanism whereby the excessively high leverage points are attributed with '0' probability and consequently having zero chances of being selected in the resampling process. The following steps represent the WLGBP method:

1. Apply the RLGD using either the MCD or the MVE to the original sample $(y_1,x_1),(y_2,x_2),\ldots,(y_n,x_n)$ and obtain $p_{ii}^{*(-D)}$.
2. Compute $u = r_{si}^{WMLE}$ by the MSPR.
3. Compute re-descending weighting function $w(u) = \psi_{Bisquare}(u)/u$.
4. Compute probability of relative weight $p_i = w_i \Big/ \sum_{i=1}^{n} w_i$.
5. The bootstrap sample $(y_1,x_1)^*,(y_2,x_2)^*,\ldots,(y_n,x_n)^*$ is taken independently with probabilities p_i from the original sample $(y_1,x_1),(y_2,x_2),\ldots,(y_n,x_n)$.
6. Compute β^* for the bootstrap sample $(y_1,x_1)^*,(y_2,x_2)^*,\ldots,(y_n,x_n)^*$.
7. Repeat Step 5 and Step 6 for B times to obtain $\hat{\beta}^{*1},\beta^{*2},\ldots,\hat{\beta}^{*B}$ bootstrap replications.

The WLGBP method is expected to be resistant to the high leverage points as the most outlying points in *X*-space are likely to receive zero probabilities and thus will have no chance of being selected.

3. Findings and Interpretation

3.1. The Simulated Data

In this section, a Monte Carlo simulation study is carried out to investigate various properties and performances of the CB, DLGBB and WLGBP methods by performing the percentile bootstrap on different percentages of high leverage points, s=(0%,5%,10%15%,20%) for n=(100,300,500) of original samples. The choice for sample size starting with n=100 is to ensure the existence and stability of the MLE. Victoria-Feser (2002) pointed out that small data may lead to unstable MLE estimates even without contamination. Following Čížek (2007), three uncontaminated x variables, are generated from a standard normal distribution, $x_p \sim N(0,1)$ with the error terms is generated according to logistic distribution, $\varepsilon_i \sim \Lambda(0,1)$. The response variable y_i is computed in the following way:

$$y_i = \begin{cases} 0 & if \quad \beta_0 + \beta_1 x_1 + \beta_2 x_2, \beta_3 x_3 + \varepsilon_i < 0 \\ 1 & if \quad \beta_0 + \beta_1 x_1 + \beta_2 x_2, \beta_3 x_3 + \varepsilon_i \geq 0 \end{cases} \tag{19}$$

Meanwhile, for contaminated data, two z variables are generated from same standard normal distribution, $z_1 \sim N(0,1)$ and $z_2 \sim N(0,1)$ with magnitude of outlying shift distance in *X* space is taken as $\delta = 10$. Then the contaminated x^* are defined as $x_1^* = z_1 + \delta$ and $x_2^* = z_2 - \delta$ with the error $\varepsilon_i \sim \Lambda(0,4)$. For the contaminated data, the response variable is computed by the following equation:

$$y_i^* = \begin{cases} 0 & if \quad \beta_0 + \beta_1 x_1^* + \beta_2 x_2^*, \beta_3 x_3 + \varepsilon_i \geq 0 \\ 1 & if \quad \beta_0 + \beta_1 x_1^* + \beta_2 x_2^*, \beta_3 x_3 + \varepsilon_i < 0 \end{cases} \tag{20}$$

We set the true parameters as $\beta^T = (\beta_0,\beta_1,\beta_2,\beta_3) = (0.5,1,-1,0)$. Then we employ the RLGD method with cut-off point, 3 for median and MAD. In order to resample uncontaminated bootstrap, $B = 1000$ bootstrap samples were drawn from n original sample. Meanwhile for contaminated bootstrap samples, $B = 1000$ bootstrap samples were drawn from $(n-s)$ good observations with s contaminated observations. 100 replications of such simulations were executed to perform percentile bootstrap to determine the percentage of times the true value of the parameter estimates are contained in the interval and the length was calculated. The same procedure is repeated for the different percentages of high leverage points. Some summary values over $B = 1000$ replications were computed, such as the mean

$$\bar{\hat{\beta}}_j^* = \frac{1}{B}\sum_{b=1}^{B} \beta_j^{*\bar{B}} \tag{21}$$

where b=1,2,…,B and j=1,2,…,p which yields the bias

$$\hat{\beta}_j^* - \beta_j^T \tag{22}$$

The estimated root mean squared error (RMSE) is given by

$$\sqrt{\left(\bar{\hat{\beta}}_j^* - \beta_j^T\right)^2 + \frac{1}{B-1}\sum_{b=1}^{B} \beta_j^{*\bar{B}} - \hat{\beta}_j^*} \tag{23}$$

In practice, we call a bootstrap method 'good' if its estimated coefficients, standard errors, biases and RMSE are reasonably closer to the MLE for uncontaminated data. A 'good' bootstrap method also should have smallest bias and RMSE in the presence of high leverage points. Another set of assessments is based on the coverage probability, the equitailness and the length of confidence interval (Habshah, 2000). By equiltailness, we mean that a confidence interval for β of level $(1-2\alpha)$ is such that proportion for β lying outside the interval is divided equally between the lower and upper limits of the intervals. In other words, the proportion of β lower than the lower limit of the interval is reasonably equal to α , as is the proportion of β that exceeds the upper limit. In this assessment, a 'good' bootstrap method is that one which have coverage probability closer to the nominal probability (95%), good equitailness and smaller average interval length. The results of the simulation studies are illustrated in Table 1-Table 11.

Let us first focus on bootstrap estimates of one contaminated covariate, x_1 for three different sample sizes (n = 100, 300 and 500) and these results are presented in Tables 1-3. We observe from these tables that when there is no high leverage point, the three bootstrap methods produce estimated coefficients, standard errors, biases and RMSE fairly close to each other and reasonably closer to the MLE estimates for clean data. Nonetheless, the MLE and the CB perform poorly in the presence of high leverage points. The estimated coefficients and standard errors for $\hat{\beta}_1$ are far from the true value $(\beta_1 = 1)$. We observe that the biases and RMSE's increase with the increase in the percentages of high leverage points. $\hat{\beta}_0$ is not much affected in terms of estimated coefficients and standard errors, but biases and RMSE's are bigger compared to those of the MLE for uncontaminated sample. As expected the DLGBB performs the best overall in terms of estimated coefficients, standard errors, biases and RMSE which are reasonably closer to the uncontaminated MLE for different levels of percentages of high leverage points. We also observe that the biases and RMSE's of the WLGBP are slightly higher than the DLGBB, but are much better than the MLE and the CB. The results are consistent for all different samples sizes considered in this study.

Next, we will focus on the results for two contaminated covariates, x_1 and x_2 as shown in Tables 4-6. We present only the results for $\hat{\beta}_1$ and $\hat{\beta}_2$ for brevity. The results of $\hat{\beta}_0$ for two contaminated covariates are similar to with one contaminated covariate. Likewise the previous results for one covariate model, the performances of CB, DLGBB and WLGBP are equally good and their estimates are fairly close to the MLE for uncontaminated data. Again the MLE and CB perform very poorly in the presence of high leverage points. Likewise the previous results the DLGBB performs best overall followed by the WLGBP for contaminated data.

Results presented in Tables 7-9 illustrate the summary statistics for three covariates in the model with two contaminated covariates. The results of $\hat{\beta}_0$ and $\hat{\beta}_1$ are consistent and are not included for brevity. Likewise the previous results, the CB, the DLGBB and the WLGBP produce estimated coefficients, standard errors, biases and RMSE's fairly close to each other and reasonably closer to the uncontaminated MLE. The performances of the MLE and the CB are not satisfactory at all in the presence of high leverage points as their estimates are far from the MLE for clean data. It is interesting to note that the DLGBB consistently estimates coefficients, standard errors, biases and RMSE's which are reasonably closer to the uncontaminated MLE for different levels of contamination followed by the WLGBP.

The findings from simulation study of estimates of coefficients, standard errors, biases and RMSE's reveal that the DLGBB outperforms other bootstrap methods in the presence of high leverage points in logistic regression model. In order to provide more evidences, the performances of CB, DLGBB and WLGBP are further investigated based on the coverage probability, equitailness and average interval length. Table 10-11 shows where the data is uncontaminated the MLE is the best estimator with the smallest average interval length, though a little over coveraged while the CB, the DLGBB and the WLGBP perform reasonably close to each other with average interval length slightly larger than uncontaminated MLE. Under contamination, the MLE and the CB give erroneous results. It can be seen from Table 10-11 that they have very bad coverage probabilities and equitailness although their average interval lengths are smaller than the DLGBB and the WLGBP. The coverage probabilities and equitailness of $\hat{\beta}_3$ for the contaminated MLE and CB are a little over coverage by 5% with wider length compared to the $\hat{\beta}_3$ for the uncontaminated MLE. As mentioned earlier, we insert high leverage points in x_1 and x_2. Therefore, $\hat{\beta}_3$ is not much affected in the presence of high leverage points. However, $\hat{\beta}_0$ gives poor coverage probabilities, especially for the contaminated MLE. On the other hand, the DLGBB and the WLGBP consistently provide good coverage probabilities.

3.2. The Real Data Sets

So far we have considered artificial data to investigate the performance of our newly proposed robust bootstrap methods. Now we apply these methods to several well-known real world data sets to investigate the performance of our newly proposed robust bootstrap methods. For assessing different estimators we consider another criterion proposed by Kordzakhia et al. (2001) where we use a Chi-square statistic based on the arcsin transformation χ^2_{arc}, defined as:

$$\chi^2_{arc} = \sum_{i=1}^{n} 4\left[\arcsin\sqrt{y_i} - \arcsin\sqrt{\pi_i}\right]^2 \qquad (24)$$

Bootstrap estimates with better fits should have lower χ^2_{arc} value**.**

3.2.1 The Modified Coronary Heart Disease Data

At first we consider the coronary heart disease data taken from Hosmer and Lemeshow (2000). This data investigated the relationship between age (Age) and the presence Y=1 or absence Y=0 of coronary heart disease (CHD) of 100 patients. The original data are free from high leverage points. In our study we contaminated this data by replacing three clean observations with three high leverage points for case 1 $(y_{11}, x_{12}) = (0,120)$, case 2 $(y_{21}, x_{22}) = (0,123)$ and case 3 $(y_{31}, x_{32}) = (0,124)$.

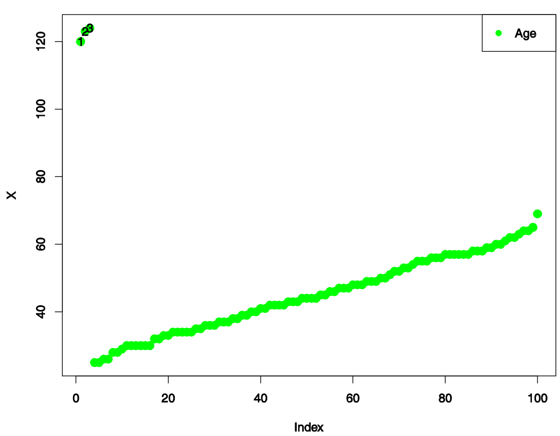

Figure 1. Index plot of Age for the modified coronary heart disease data.

The index plot of Age as shown in Figure 1 clearly shows the existence of three *X*-outliers (cases 1, 2, and 3) in this data.

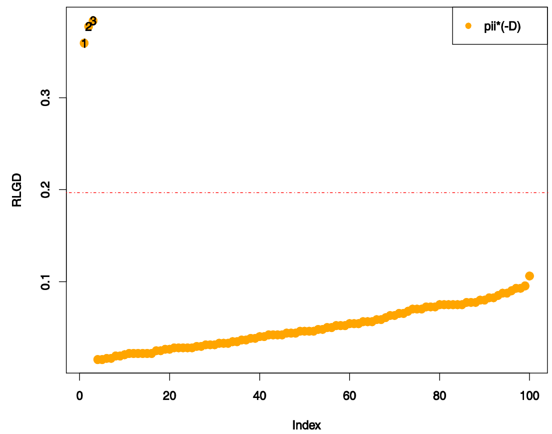

Figure 2. Index plot of RLGD for the modified coronary heart disease data.

For confirmation we apply the RLGD method based on the MCD which yields the cut-off point 0.1968. We compute the group deleted potentials for the entire data set and observe that the RLGD values corresponding to cases 1, 2 and 3 which are 0.3593, 0.3775, and 0.3837 respectively are much higher than the cut-off point, which reveal that these three cases are high leverage points. Similar conclusion may be drawn from the index plot of RLGD as shown in Figure 2. All the three suspected cases are clearly separated from the rest of the data.

Table 12 presents estimates of parameters, standard errors and goodness of fit tests for the modified coronary heart disease data. The MLE_97 in Table 1 refers to the ML estimates for the clean data after the three inserted high leverage points are omitted from the data. The MLE_100 are the ML estimates for 100 observations including the three high leverage points. We observe from Table 1 that the MLE_100 gives the worst set of results in term of estimated coefficient and standard error for Age which is far from the MLE_97 and gives the highest χ^2_{arc}. The CB is also severely affected by the high leverage points. But the newly proposed DLGBB performs best overall. The estimated coefficients and standard errors of the DLGBB are very close to those of the MLE_97 and this method yields the smallest χ^2_{arc} value. Even though the WLGBP gives the highest standard errors, its χ^2_{arc} value is smaller compared to the MLE and the CB.

3.2.2. The Prostate Cancer Data

Our next example is the prostate cancer data given by Brown (1980). Here the objective was to see whether two continuous variables, an elevated level of acid phosphates in the blood serum, (AP) and age of 53 patients (Age) would be of value for predicting whether or not prostate cancer patients also had lymph node involvement (LNI).

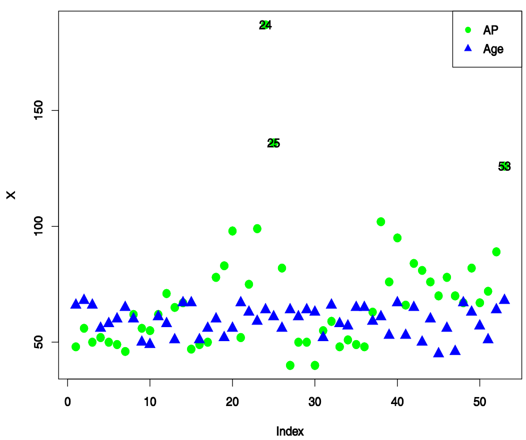

Figure 3. Index plot of AP and Age for the prostate cancer data.

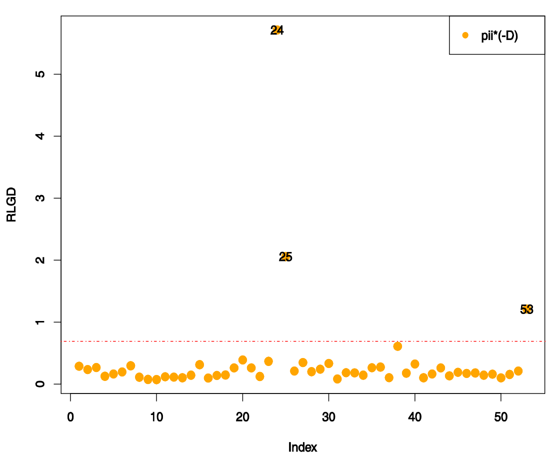

Figure 4. Index plot of RLGD for the prostate cancer data.

It was reported by many authors (Hadi, 1992; Ryan, 1997; Imon, 2006) that there exist three leverage points (cases 24, 25 and 53) in this data. The index plot of AP and Age illustrates these three high leverage points. These three observations are confirmed as high leverage points when the RLGD method is employed. The RLGD values for these three observations are 5.7149, 2.0575 and 1.2104 respectively while the cut-off point is 0.6900. The index plot of the RLGD values as shown in Figure 4 reconfirms the above statement.

We observe from the results presented in Table 2 that the MLE_53 and the CB give estimated coefficients which are far from the MLE_50 and they also possess higher χ^2_{arc}. The DLGBB performs the best followed by the WLGBP with estimated coefficients which are close to the MLE_50 and give the smallest χ^2_{arc} value. On deleting the high leverage points, the remaining data may have less overlapping cases. This may be the reason for which the DLGBP and the WLGBP possessing a bit higher standard errors.

3.2.3. The Intensive Care Unit Data

Our final example is the intensive care unit (ICU) data which consists of 200 subjects related to survival of patients following admission to an adult intensive care unit. The major goal of this study was to predict the probability of survival status (STA) to hospital discharge of cancer patients and to study three risk factors of age at ICU admission (Age), systolic blood pressure at ICU admission (SYS) in mm/Hg and heart rate at ICU admission (HRA) in Beat/min which are associated with ICU mortality. Data were collected at Baystate medical center in Springfield, Massachusetts and provided by Hosmer and Lemeshow (2000).

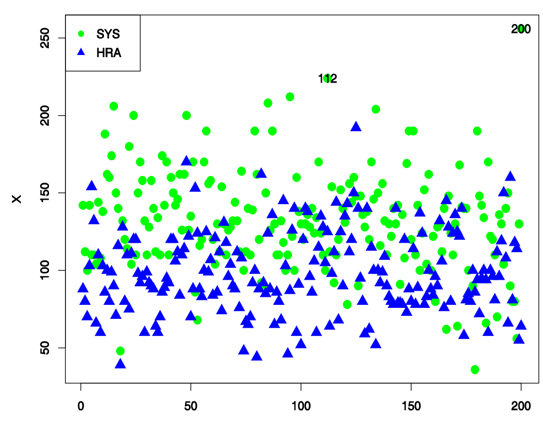

Figure 5. Index plot of SYS and HRA for the intensive care unit data.

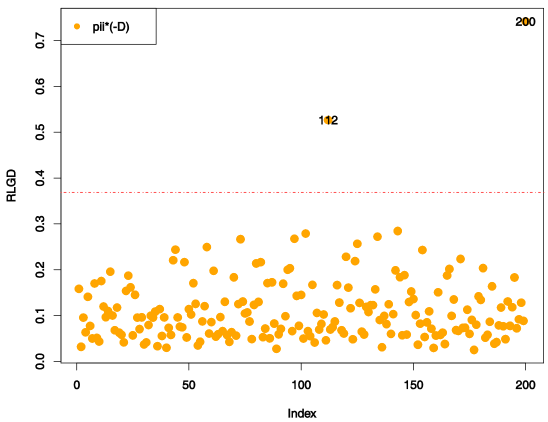

Figure 6. Index plot of RLGD for the intensive care unit data.

The index plot of SYS and HRA clearly suggests that observations 112 and 200 may be outlying in the covariate space of SYS. For the confirmation we employ the RLGD method here. The RLGD values corresponding to cases 112 and 200 are 0.5266 and 0.7417 respectively which exceed the cut-off point 0.3690. Thus these two observations are confirmed as high leverage points. The index plot of RLGD as shown in Figure 6 also supports this statement.

Results presented in Table 3 give exactly same kind of conclusions as we got in the previous two examples. The MLE_200 and the CB produce the worst results. Their χ^2_{arc} values are large. The DLGBB and the WLGBP give satisfactory results where their parameter estimates are reasonably closer to the MLE_198 and have smaller χ^2_{arc} compared to the MLE_200 and the CB.

4. Discussions

In this paper we propose two techniques for bootstrapping logistic regression model in the presence of high leverage points. The simulation studies suggest that for clean data, the classical bootstrap, and the proposed DLGBB and the WLGBP performs similarly in terms of coverage probabilities, equitailness and average interval length. But the CB suffers a huge set back in the presence of high leverage points. In this situation, the CB produces remarkably low coverage probabilities, poor equitailness and wider average interval length. However, the DLGBB and the WLGBP consistently provide adequate coverage probabilities, good equitailness and shorter average interval length. The real world examples show that the DLGBB and the WLGBP provide smaller values of goodness of fit, χ^2_{arc} in comparison with the CB. The results consistently show that the DLGBB is performs better than the WLGBP. Hence, the DLGBB method can be considered as the best techniques for bootstrapping logistic regression model in the presence of high leverage points.

Acknowledgements:

Foundation item: The Fundamental Research Grant Scheme (Vote No.: 5524075). Authors are grateful to the Ministry of Higher Education of Malaysia for financial support to carry out this work.

Corresponding Author:
Syaiba Balqish Ariffin
Department of Mathematics
Faculty of Science, Universiti Putra Malaysia
43400 Serdang, Selangor, Malaysia
E-mail: syaibabalqish@gmail.com

References

[1]. Amado C, Pires AM. Robust bootstrap with non-random weights based on the influence function. Communications in Statistics, Simulation and Computation 2004;33(2):377-396.
[2]. Brown BW. Prediction Analysis for Binary Data. In Biostatistics Casebook. R. G. Miller, B. Efron, B. W. Brown and L. E. Moses (editors). Wiley: New York, 1980: 3-18.
[3]. Čížek P. Robust and efficient adaptive estimation of binary-choice regression models. CentER Discussion Paper, Tilburg University 2007, 12:1-30.
[4]. Croux C, Flandre C, Haesbroeck G. The breakdown behaviour of the maximum likelihood estimator in the logistic regression model. Statistics & Probability Letter 2002;60(4):377-386.
[5]. Croux C, Haesbroeck G. Implementing the Bianco and Yohai estimator for logistic regression. Computational Statistics & Data Analysis 2003;44(1-2):273-295.
[6]. Efron B. Bootstrap method: Another look at the Jackknife. The Annals of Statistics 1979;7(1):1-26.
[7]. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. Chapman and Hall/CRC: Boca Raton, 1998: 105-120.
[8]. Imon AHMR, Ali MM. Bootstrapping regression residual. Journal of Korean Data and Information Science Society 2005;16(3):665-682.

[9]. Habshah M. Bootstrap methods: A class of non-linear regression models. Pertanika Journal of Science and Technology 2000;8(2):175-189.
[10]. Habshah M, Hassan SU, Bashar AT. Dynamic robust bootstrap method based on LTS estimators. European Journal of Scientific Research 2009;32(3):277-287.
[11]. Hadi AS. A new measure of overall potential influence in linear regression. Journal of Computational Statistics & Data Analysis 1992;14(1):1-27.
[12]. Hosmer DW, Lemeshow S. Applied Logistic Regression, 2nd ed. Wiley:New York, 2000:1-30.
[13]. Hossain A, Khan HTA. Nonparametric bootstrapping for multiple logistic regression model using R. BRAC University Journal 2004;1(2):109-113.
[14]. Hubert M, Rousseeuw PJ. Robust regression with both continuous and binary regresors. Journal of Statistical Planning and Inference 1997;57(1):153-163.
[15]. Imon AHMR. Identification of high leverage points in logistic regression model. Pakistan Journal Statistics 2006;22(2):147-156.
[16]. Izrael D, Battaglia AA, Hoaglin DC, Battaglia MP. Use of the ROC curve and the bootstrap in computing weighted logistic regression models. In Proceedings of the Twenty-seventh Annual SAS Users Group International Conference, Cary, Sugi27, Statistics and Data Analysis, 2002:248-27.
[17]. Kordzakhia N, Mishra GD, Reiersølmoen L. Robust estimation in logistic regression model. Journal of Statistical Planning and Inference 2001;98(1-2):211-223.
[18]. Norazan MR. Weighted Maximum Median Likelihood Estimation for Parameters in Multiple Linear Regression Model. PhD Thesis, Universiti Putra Malaysia, 2008: 269-332.
[19]. Norazan MR, Habshah M, Imon AHMR. Estimating regression coefficients using weighted bootstrap with probability. WSEAS Transactions on Mathematics 2009;8(7):362-371.
[20]. Roberts G, Binder D, Kovačević M, Pantel M, Phillips O. Using an estimating function bootstrap approach for obtaining variances estimates when modelling complex health survey data. In Proceeding of the Survey Methods Sections, SSC Annual Meeting, 2003: 1-9.
[21]. Rousseeuw PJ, Leroy AM. Robust Regression and Outlier Detection. Wiley: New York, 1987:248-273.
[22]. Ryan TP. Modern Regression Methods. Wiley:New York, 1997:275-300.
[23]. Salibian-Barrera M. Bootstrapping MM-estimators for linear regression with fixed designs. Statistics & Probability Letter 2006;76(12):1287-1297.
[24]. Salibian-Barrera M, Zamar RH. Bootstrapping robust estimates of regression. The Annals of Statistics 2002;30(2):556-582.
[25]. Shao J. Bootstrap variance estimators with truncation. Statistics and Probability Letters 1992;15(2):95-101.
[26]. Stromberg AJ. Robust covariance estimates based on resampling. Journal of Statistical Planning and Inference 1997;57(2):321-334.
[27]. Syaiba BA, Habshah M. Robust logistic diagnostic for the identification of high leverage points in logistic regression model. Journal of Applied Sciences 2010;10(23):3042-3050.
[28]. Victoria-Feser M-P. Robust inference with binary data. Psychometrika 2002; 67(1):21-32.
[29]. Willems G, Aelst SV. Fast and robust bootstrap for LTS. Computational Statistics & Data Analysis 2005;48(4):703-715.

Table 1. Bootstrap estimates for one covariate with $n = 100$.

% HLP	Estimation Methods	$\hat{\beta}_0$ Value	Std.Err	Bias	RMSE	$\hat{\beta}_1$ Value	Std.Err	Bias	RMSE
0	MLE	0.5102	0.2402	0.0102	0.2404	1.0610	0.2938	0.0610	0.3001
	CB	0.5260	0.3556	0.0260	0.3565	1.1222	0.4281	0.1222	0.4452
	DLGBB	0.5260	0.3556	0.0260	0.3565	1.1222	0.4281	0.1222	0.4452
	WLGBP	0.5260	0.3556	0.0260	0.3565	1.1222	0.4281	0.1222	0.4452
5	MLE	0.3290	0.2156	-0.1710	0.2752	-0.0492	0.0574	-1.0492	1.0508
	CB	0.3301	0.3001	-0.1699	0.3449	0.0001	0.2065	-0.9999	1.0210
	DLGBB	0.5169	0.3482	0.0169	0.3486	1.0914	0.4579	0.0914	0.4669
	WLGBP	0.6788	0.4414	0.1788	0.4762	1.4189	0.6754	0.4188	0.7947
10	MLE	0.3166	0.2234	-0.1834	0.2890	-0.1305	0.0500	-1.1305	1.1316
	CB	0.3154	0.3183	-0.1846	0.3680	-0.1272	0.0802	-1.1272	1.1301
	DLGBB	0.5344	0.3688	0.0344	0.3704	1.1104	0.4819	0.1104	0.4944
	WLGBP	0.6855	0.4575	0.1855	0.4937	1.4200	0.6864	0.4200	0.8047
15	MLE	0.3153	0.2340	-0.1847	0.2981	-0.1761	0.0501	-1.1761	1.1772
	CB	0.3172	0.3317	-0.1828	0.3788	-0.1775	0.0741	-1.1775	1.1799
	DLGBB	0.5494	0.3878	0.0494	0.3909	1.0960	0.4773	0.0960	0.4868
	WLGBP	0.6736	0.4660	0.1736	0.4973	1.3818	0.7545	0.3818	0.8456
20	MLE	0.3204	0.2341	-0.1796	0.2950	-0.2006	0.0500	-1.2006	1.2017
	CB	0.3204	0.3250	-0.1796	0.3713	-0.2053	0.0727	-1.2053	1.2075
	DLGBB	0.5267	0.3863	0.0267	0.3872	1.0790	0.4837	0.0790	0.4901
	WLGBP	0.6895	0.4792	0.1895	0.5153	1.3460	0.7728	0.3460	0.8467

Table 2. Bootstrap estimates for one covariate with $n = 300$.

% HLP	Estimation Methods	$\hat{\beta}_0$ Value	Std.Err	Bias	RMSE	$\hat{\beta}_1$ Value	Std.Err	Bias	RMSE
0	MLE	0.5068	0.1358	0.0068	0.1360	1.0198	0.1563	0.0198	0.1576
	CB	0.5086	0.1905	0.0086	0.1907	1.0360	0.2253	0.0360	0.2281
	DLGBB	0.5086	0.1905	0.0086	0.1907	1.0360	0.2253	0.0360	0.2281
	WLGBP	0.5086	0.1905	0.0086	0.1907	1.0360	0.2253	0.0360	0.2281
5	MLE	0.3281	0.1171	-0.1719	0.2080	-0.0460	0.0319	-1.0460	1.0465
	CB	0.3245	0.1709	-0.1755	0.2450	-0.0375	0.0675	-1.0375	1.0397
	DLGBB	0.5068	0.1895	0.0068	0.1896	1.0232	0.2444	0.0232	0.2455
	WLGBP	0.6533	0.2344	0.1533	0.2801	1.3495	0.2988	0.3495	0.4599
10	MLE	0.3237	0.1275	-0.1763	0.2176	-0.1306	0.0281	-1.1306	1.1309
	CB	0.3222	0.1828	-0.1778	0.2551	-0.1294	0.0443	-1.1294	1.1303
	DLGBB	0.5087	0.2047	0.0087	0.2048	1.0369	0.2454	0.0369	0.2482
	WLGBP	0.6661	0.2437	0.1661	0.2950	1.3449	0.3072	0.3449	0.4619
15	MLE	0.3195	0.1336	-0.1805	0.2246	-0.1705	0.0280	-1.1705	1.1708
	CB	0.3256	0.1932	-0.1744	0.2603	-0.1722	0.0400	-1.1722	1.1729
	DLGBB	0.5044	0.2090	0.0044	0.2091	1.0389	0.2599	0.0388	0.2628
	WLGBP	0.6617	0.2319	0.1617	0.2827	1.3494	0.3170	0.3494	0.4718
20	MLE	0.3134	0.1367	-0.1866	0.2313	-0.1944	0.0287	-1.1944	1.1947
	CB	0.3167	0.1952	-0.1833	0.2678	-0.1973	0.0410	-1.1973	1.1980
	DLGBB	0.5125	0.2171	0.0125	0.2174	1.0338	0.2594	0.0338	0.2616
	WLGBP	0.6603	0.2529	0.1603	0.2995	1.3478	0.3266	0.3478	0.4772

Table 3. Bootstrap estimates for one covariate with $n = 500$.

% HLP	Estimation Methods	$\hat{\beta}_0$ Value	$\hat{\beta}_0$ Std.Err	$\hat{\beta}_0$ Bias	$\hat{\beta}_0$ RMSE	$\hat{\beta}_1$ Value	$\hat{\beta}_1$ Std.Err	$\hat{\beta}_1$ Bias	$\hat{\beta}_1$ RMSE
0	MLE	0.5052	0.1040	0.0052	0.1041	1.0004	0.1157	0.0004	0.1157
	CB	0.5165	0.1490	0.0165	0.1499	1.0156	0.1740	0.0156	0.1747
	DLGBB	0.5165	0.1490	0.0165	0.1499	1.0156	0.1740	0.0156	0.1747
	WLGBP	0.5165	0.1490	0.0165	0.1499	1.0156	0.1740	0.0156	0.1747
5	MLE	0.3286	0.0906	-0.1714	0.1939	-0.0471	0.0235	-1.0471	1.0473
	CB	0.3320	0.1262	-0.1680	0.2101	-0.0418	0.0469	-1.0418	1.0429
	DLGBB	0.5096	0.1464	0.0096	0.1467	1.0160	0.1869	0.0160	0.1876
	WLGBP	0.6537	0.1703	0.1537	0.2295	1.3192	0.2343	0.3192	0.3959
10	MLE	0.3250	0.0980	-0.1750	0.2006	-0.1299	0.0213	-1.1299	1.1301
	CB	0.3287	0.1392	-0.1713	0.2207	-0.1298	0.0340	-1.1298	1.1303
	DLGBB	0.5094	0.1546	0.0094	0.1549	1.0209	0.1884	0.0209	0.1896
	WLGBP	0.6630	0.1802	0.1630	0.2430	1.3235	0.2383	0.3235	0.4018
15	MLE	0.3211	0.1016	-0.1789	0.2058	-0.1688	0.0213	-1.1688	1.1690
	CB	0.3293	0.1481	-0.1707	0.2260	-0.1700	0.0314	-1.1700	1.1704
	DLGBB	0.5081	0.1613	0.0081	0.1615	1.0199	0.1876	0.0199	0.1887
	WLGBP	0.6568	0.1865	0.1568	0.2436	1.3200	0.2358	0.3200	0.3975
20	MLE	0.3160	0.1034	-0.1840	0.2110	-0.1936	0.0211	-1.1936	1.1937
	CB	0.3165	0.1446	-0.1835	0.2337	-0.1942	0.0315	-1.1942	1.1946
	DLGBB	0.5028	0.1643	0.0028	0.1643	1.0234	0.1937	0.0234	0.1951
	WLGBP	0.6560	0.1845	0.1560	0.2416	1.3296	0.2431	0.3296	0.4095

Table 4. Bootstrap estimates for two covariates with $n = 100$.

% HLP	Estimation Methods	$\hat{\beta}_1$ Value	$\hat{\beta}_1$ Std.Err	$\hat{\beta}_1$ Bias	$\hat{\beta}_1$ RMSE	$\hat{\beta}_2$ Value	$\hat{\beta}_2$ Std.Err	$\hat{\beta}_2$ Bias	$\hat{\beta}_2$ RMSE
0	MLE	1.0670	0.3057	0.0670	0.3130	-1.0440	0.3074	-0.0440	0.3106
	CB	1.1291	0.4686	0.1291	0.4861	-1.1038	0.4688	-0.1038	0.4801
	DLGBB	1.1291	0.4686	0.1291	0.4861	-1.1038	0.4688	-0.1038	0.4801
	WLGBP	1.1291	0.4686	0.1291	0.4861	-1.1038	0.4688	-0.1038	0.4801
5	MLE	-0.0454	0.1567	-1.0454	1.0571	0.0433	0.1532	1.0433	1.0545
	CB	-0.0154	0.2565	-1.0154	1.0473	0.0160	0.2744	1.0160	1.0524
	DLGBB	1.1305	0.5121	0.1305	0.5285	-1.1333	0.4940	-0.1333	0.5116
	WLGBP	1.5627	0.7118	0.5627	0.9073	-1.5687	0.7106	-0.5687	0.9102
10	MLE	-0.0919	0.1540	-1.0919	1.1027	0.0891	0.1536	1.0891	1.0999
	CB	-0.0939	0.2382	-1.0939	1.1195	0.0837	0.2356	1.0837	1.1091
	DLGBB	1.1406	0.5314	0.1406	0.5496	-1.1623	0.5408	-0.1623	0.5646
	WLGBP	1.5631	0.7249	0.5631	0.9179	-1.5901	0.6894	-0.5901	0.9075
15	MLE	-0.1161	0.1629	-1.1161	1.1279	0.1166	0.1643	1.1166	1.1287
	CB	-0.1156	0.2419	-1.1156	1.1415	0.1189	0.2442	1.1189	1.1452
	DLGBB	1.1296	0.5286	0.1296	0.5442	-1.1529	0.5618	-0.1529	0.5822
	WLGBP	1.5678	0.8513	0.5678	1.0233	-1.5975	0.9722	-0.5975	1.1412
20	MLE	-0.1339	0.1667	-1.1339	1.1461	0.1335	0.1651	1.1335	1.1455
	CB	-0.1361	0.2385	-1.1361	1.1609	0.1390	0.2383	1.1390	1.1637
	DLGBB	1.1647	0.5772	0.1647	0.6002	-1.1726	0.5890	-0.1726	0.6138
	WLGBP	1.5559	0.7547	0.5559	0.9373	-1.6045	0.8606	-0.6045	1.0516

Table 5. Bootstrap estimates for two covariates with $n = 300$.

% HLP	Estimation Methods	$\hat{\beta}_1$ Value	Std.Err	Bias	RMSE	$\hat{\beta}_2$ Value	Std.Err	Bias	RMSE
0	MLE	1.0264	0.1735	0.0264	0.1755	-1.0227	0.1761	-0.0227	0.1775
	CB	1.0443	0.2465	0.0443	0.2505	-1.0437	0.2410	-0.0437	0.2449
	DLGBB	1.0443	0.2465	0.0443	0.2505	-1.0437	0.2410	-0.0437	0.2449
	WLGBP	1.0443	0.2465	0.0443	0.2505	-1.0437	0.2410	-0.0437	0.2449
5	MLE	-0.0444	0.0847	-1.0444	1.0479	0.0428	0.0830	1.0428	1.0461
	CB	-0.0453	0.1264	-1.0453	1.0529	0.0349	0.1265	1.0349	1.0426
	DLGBB	1.0289	0.2500	0.0289	0.2517	-1.0407	0.2545	-0.0407	0.2578
	WLGBP	1.3915	0.3441	0.3915	0.5212	-1.3960	0.3337	-0.3960	0.5178
10	MLE	-0.0862	0.0861	-1.0862	1.0896	0.0925	0.0858	1.0925	1.0959
	CB	-0.0875	0.1246	-1.0875	1.0946	0.0911	0.1226	1.0911	1.0979
	DLGBB	1.0460	0.2559	0.0460	0.2600	-1.0427	0.2512	-0.0427	0.2548
	WLGBP	1.4199	0.3423	0.4199	0.5418	-1.4062	0.3443	-0.4062	0.5325
15	MLE	-0.1103	0.0874	-1.1103	1.1138	0.1179	0.0880	1.1179	1.1213
	CB	-0.1115	0.1270	-1.1115	1.1188	0.1173	0.1266	1.1173	1.1245
	DLGBB	1.0530	0.2701	0.0530	0.2753	-1.0496	0.2681	-0.0496	0.2726
	WLGBP	1.4088	0.3552	0.4088	0.5416	-1.4130	0.3468	-0.4130	0.5393
20	MLE	-0.1346	0.0953	-1.1346	1.1386	0.1278	0.0948	1.1278	1.1318
	CB	-0.1358	0.1314	-1.1358	1.1434	0.1288	0.1319	1.1288	1.1365
	DLGBB	1.0393	0.2729	0.0393	0.2758	-1.0440	0.2801	-0.0440	0.2836
	WLGBP	1.4012	0.3571	0.4012	0.5371	-1.4203	0.3688	-0.4203	0.5591

Table 6. Bootstrap estimates for two covariates with $n = 500$.

% HLP	Estimation Methods	$\hat{\beta}_1$ Value	Std.Err	Bias	RMSE	$\hat{\beta}_2$ Value	Std.Err	Bias	RMSE
0	MLE	1.0028	0.1235	0.0028	0.1235	-1.0103	0.1299	-0.0103	0.1303
	CB	1.0124	0.1814	0.0124	0.1819	-1.0161	0.1857	-0.0161	0.1864
	DLGBB	1.0124	0.1814	0.0124	0.1819	-1.0161	0.1857	-0.0161	0.1864
	WLGBP	1.0124	0.1814	0.0124	0.1819	-1.0161	0.1857	-0.0161	0.1864
5	MLE	-0.0417	0.0646	-1.0417	1.0437	0.0447	0.0655	1.0447	1.0467
	CB	-0.0384	0.0967	-1.0384	1.0429	0.0432	0.0966	1.0432	1.0476
	DLGBB	1.0183	0.1908	0.0183	0.1917	-1.0161	0.1945	-0.0161	0.1951
	WLGBP	1.3696	0.2474	0.3696	0.4447	-1.3625	0.2478	-0.3625	0.4391
10	MLE	-0.0826	0.0678	-1.0826	1.0847	0.0957	0.0673	1.0957	1.0978
	CB	-0.0828	0.0977	-1.0828	1.0872	0.0957	0.0983	1.0957	1.1001
	DLGBB	1.0297	0.2004	0.0297	0.2026	-1.0136	0.1890	-0.0136	0.1895
	WLGBP	1.3815	0.2588	0.3815	0.4610	-1.3603	0.2427	-0.3603	0.4344
15	MLE	-0.1138	0.0667	-1.1138	1.1158	0.1134	0.0669	1.1134	1.1155
	CB	-0.1171	0.0988	-1.1171	1.1214	0.1104	0.0992	1.1104	1.1148
	DLGBB	1.0289	0.1917	0.0289	0.1939	-1.0342	0.2035	-0.0342	0.2063
	WLGBP	1.3859	0.2605	0.3859	0.4656	-1.3804	0.2562	-0.3804	0.4587
20	MLE	-0.1273	0.0676	-1.1273	1.1293	0.1345	0.0684	1.1345	1.1365
	CB	-0.1305	0.0991	-1.1305	1.1349	0.1323	0.1013	1.1323	1.1369
	DLGBB	1.0316	0.2120	0.0316	0.2144	-1.0319	0.2140	-0.0319	0.2163
	WLGBP	1.3879	0.2592	0.3879	0.4665	-1.3746	0.2704	-0.3746	0.4620

Table 7. Bootstrap estimates for three covariates with $n = 100$.

%	Estimation	$\hat{\beta}_2$				$\hat{\beta}_3$			
HLP	Methods	Value	Std.Err	Bias	RMSE	Value	Std.Err	Bias	RMSE
0	MLE	-1.0588	0.3173	-0.0588	0.3227	0.0054	0.2832	0.0054	0.2833
	CB	-1.1538	0.5082	-0.1538	0.5309	0.0130	0.4259	0.0130	0.4261
	DLGBB	-1.1538	0.5082	-0.1538	0.5309	0.0130	0.4259	0.0130	0.4261
	WLGBP	-1.1538	0.5082	-0.1538	0.5309	0.0130	0.4259	0.0130	0.4261
5	MLE	0.0432	0.1583	1.0432	1.0552	0.0022	0.2277	0.0022	0.2277
	CB	0.0105	0.2701	1.0105	1.0460	0.0014	0.3392	0.0014	0.3392
	DLGBB	-1.1500	0.5421	-0.1500	0.5625	0.0177	0.4425	0.0177	0.4428
	WLGBP	-1.6569	0.8123	-0.6569	1.0446	0.0019	0.5640	0.0019	0.5640
10	MLE	0.0888	0.1570	1.0888	1.1000	0.0062	0.2193	0.0062	0.2194
	CB	0.0993	0.2395	1.0993	1.1251	0.0015	0.3301	0.0015	0.3301
	DLGBB	-1.1933	0.5594	-0.1933	0.5918	-0.0185	0.4479	-0.0185	0.4482
	WLGBP	-1.6917	0.8949	-0.6917	1.1311	0.0074	0.5556	0.0074	0.5557
15	MLE	0.1171	0.1666	1.1171	1.1295	0.0022	0.2213	0.0022	0.2213
	CB	0.1180	0.2565	1.1180	1.1470	-0.0085	0.3400	-0.0085	0.3401
	DLGBB	-1.2152	0.6602	-0.2152	0.6944	-0.0209	0.4738	-0.0209	0.4743
	WLGBP	-1.6681	0.8140	-0.6681	1.0531	-0.0172	0.5528	-0.0172	0.5531
20	MLE	0.1357	0.1673	1.1357	1.1479	0.0043	0.2304	0.0043	0.2305
	CB	0.1471	0.2539	1.1471	1.1748	-0.0004	0.3490	-0.0004	0.3490
	DLGBB	-1.1965	0.5865	-0.1965	0.6185	0.0064	0.4858	0.0064	0.4858
	WLGBP	-1.7170	1.0910	-0.7170	1.3055	-0.0181	0.6345	-0.0181	0.6347

Table 8. Bootstrap estimates for three covariates with $n = 300$.

%	Estimation	$\hat{\beta}_2$				$\hat{\beta}_3$			
HLP	Methods	Value	Std.Err	Bias	RMSE	Value	Std.Err	Bias	RMSE
0	MLE	-1.0262	0.1773	-0.0262	0.1792	0.0057	0.1367	0.0057	0.1369
	CB	-1.0495	0.2471	-0.0495	0.2520	0.0096	0.2036	0.0096	0.2038
	DLGBB	-1.0495	0.2471	-0.0495	0.2520	0.0096	0.2036	0.0096	0.2038
	WLGBP	-1.0495	0.2471	-0.0495	0.2520	0.0096	0.2036	0.0096	0.2038
5	MLE	0.0423	0.0832	1.0423	1.0456	0.0016	0.1175	0.0016	0.1176
	CB	0.0370	0.1208	1.0370	1.0440	-0.0035	0.1789	-0.0035	0.1790
	DLGBB	-1.0492	0.2545	-0.0492	0.2593	0.0064	0.2173	0.0064	0.2174
	WLGBP	-1.4130	0.3385	-0.4130	0.5340	0.0126	0.2633	0.0126	0.2636
10	MLE	0.0931	0.0865	1.0931	1.0965	0.0007	0.1255	0.0007	0.1255
	CB	0.0938	0.1278	1.0938	1.1013	-0.0057	0.1785	-0.0057	0.1786
	DLGBB	-1.0449	0.2651	-0.0449	0.2689	0.0131	0.2224	0.0131	0.2228
	WLGBP	-1.4318	0.3568	-0.4318	0.5601	0.0205	0.2637	0.0205	0.2645
15	MLE	0.1174	0.0875	1.1174	1.1208	-0.0045	0.1280	-0.0045	0.1281
	CB	0.1153	0.1272	1.1153	1.1226	-0.0060	0.1864	-0.0060	0.1865
	DLGBB	-1.0553	0.2734	-0.0553	0.2789	0.0013	0.2340	0.0013	0.2340
	WLGBP	-1.4243	0.3448	-0.4243	0.5467	-0.0142	0.2716	-0.0142	0.2719
20	MLE	0.1285	0.0946	1.1285	1.1325	-0.0019	0.1337	-0.0019	0.1337
	CB	0.1289	0.1319	1.1289	1.1366	0.0026	0.1907	0.0026	0.1907
	DLGBB	-1.0732	0.2840	-0.0732	0.2933	-0.0067	0.2328	-0.0067	0.2329
	WLGBP	-1.4336	0.3642	-0.4336	0.5662	-0.0114	0.2763	-0.0114	0.2765

Table 9. Bootstrap estimates for three covariates with $n = 500$.

% HLP	Estimation Method	$\hat{\beta}_2$ Value	Std.Err	Bias	RMSE	$\hat{\beta}_3$ Value	Std.Err	Bias	RMSE
0	MLE	-1.0122	0.1301	-0.0122	0.1306	0.0028	0.1046	0.0028	0.1046
	CB	-1.0244	0.1879	-0.0244	0.1895	0.0000	0.1470	0.0000	0.1470
	DLGBB	-1.0244	0.1879	-0.0244	0.1895	0.0000	0.1470	0.0000	0.1470
	WLGBP	-1.0244	0.1879	-0.0244	0.1895	0.0000	0.1470	0.0000	0.1470
5	MLE	0.0453	0.0660	1.0453	1.0473	0.0022	0.0930	0.0022	0.0930
	CB	0.0425	0.0955	1.0425	1.0469	0.0039	0.1333	0.0039	0.1333
	DLGBB	-1.0204	0.1871	-0.0204	0.1882	0.0014	0.1617	0.0014	0.1618
	WLGBP	-1.3733	0.2556	-0.3733	0.4524	0.0006	0.1965	0.0006	0.1965
10	MLE	0.0963	0.0680	1.0963	1.0984	0.0033	0.0955	0.0033	0.0956
	CB	0.0941	0.0967	1.0941	1.0984	0.0055	0.1374	0.0055	0.1375
	DLGBB	-1.0194	0.1980	-0.0194	0.1989	-0.0006	0.1719	-0.0006	0.1719
	WLGBP	-1.3813	0.2551	-0.3813	0.4588	0.0019	0.2056	0.0019	0.2056
15	MLE	0.1137	0.0670	1.1137	1.1157	-0.0008	0.0997	-0.0008	0.0997
	CB	0.1167	0.0965	1.1167	1.1208	0.0006	0.1397	0.0006	0.1397
	DLGBB	-1.0242	0.2011	-0.0242	0.2025	0.0045	0.1701	0.0045	0.1701
	WLGBP	-1.3857	0.2554	-0.3857	0.4626	-0.0030	0.2034	-0.0030	0.2034
20	MLE	0.1350	0.0687	1.1350	1.1371	0.0003	0.1023	0.0003	0.1023
	CB	0.1295	0.0983	1.1295	1.1337	-0.0011	0.1407	-0.0011	0.1407
	DLGBB	-1.0335	0.2130	-0.0335	0.2156	0.0044	0.1834	0.0044	0.1834
	WLGBP	-1.3844	0.2762	-0.3844	0.4733	0.0072	0.2090	0.0072	0.2091

Table 10. Lower coverage (LC), upper coverage (UC), coverage probability (CP), confidence interval (CI) and average interval length (Length) of $\hat{\beta}_0$ and $\hat{\beta}_1$ on bootstrap procedures for $n = 500$.

% HLP	Estimation Methods	$\hat{\beta}_0$				$\hat{\beta}_1$			
		LC	CP	UC	(CI), Length	LC	CP	UC	(CI), Length
0	MLE	0	100	0	(0.2941 - 0.7261) 0.4320	0	100	0	(0.7743 - 1.2762) 0.5019
	CB	0	100	0	(0.2098 - 0.8268) 0.6170	0	100	0	(0.6891 - 1.4092) 0.7201
	DLGBB	0	100	0	(0.2098 - 0.8268) 0.6170	0	100	0	(0.6891 - 1.4092) 0.7201
	WLGBP	0	100	0	(0.2098 - 0.8268) 0.6170	0	100	0	(0.6891 - 1.4092) 0.7201
5	MLE	0	0	100	(0.1052 - 0.4757) 0.3706	0	0	100	(-0.1722 - 0.0839) 0.2561
	CB	0	100	0	(0.0312 - 0.5606) 0.5295	0	0	100	(-0.2285 - 0.1455) 0.3740
	DLGBB	0	100	0	(0.2008 - 0.8375) 0.6367	0	100	0	(0.6759 - 1.4297) 0.7538
	WLGBP	0	100	0	(0.3053 - 1.1032) 0.7979	0	100	0	(0.9307 - 1.9282) 0.9975
10	MLE	0	2	98	(0.0909 - 0.4829) 0.3920	0	0	100	(-0.2207 - 0.0417) 0.2625
	CB	0	100	0	(0.0135 - 0.5701) 0.5566	0	0	100	(-0.2782 - 0.0976) 0.3760
	DLGBB	0	100	0	(0.1931 - 0.8511) 0.6580	0	100	0	(0.6678 - 1.4404) 0.7727
	WLGBP	0	100	0	(0.2982 - 1.1140) 0.8158	0	100	0	(0.9252 - 1.9379) 1.0126
15	MLE	0	8	92	(0.0839 - 0.4889) 0.4050	0	0	100	(-0.2498 - 0.0207) 0.2704
	CB	0	100	0	(-0.0009 - 0.5769) 0.5779	0	0	100	(-0.3093 - 0.0772) 0.3865
	DLGBB	0	100	0	(0.1867 - 0.8619) 0.6753	0	100	0	(0.6617 - 1.453941) 0.7923
	WLGBP	0	100	0	(0.2931- 1.1220) 0.8289	0	100	0	(0.9202 - 1.9483) 1.0281
20	MLE	0	28	72	(0.0727 - 0.4935) 0.4208	0	0	100	(-0.2708 - 0.0070) 0.2778
	CB	0	100	0	(-0.0131 - 0.5923) 0.6054	0	0	100	(-0.3325 - 0.0659) 0.3984
	DLGBB	0	100	0	(0.1749 - 0.8745) 0.6997	0	100	0	(0.6522 - 1.4704) 0.8183
	WLGBP	0	100	0	(0.2853 -1.1334) 0.8481	0	100	0	(0.9112 - 1.9597) 1.0485

Table 11. Lower coverage (LC), upper coverage (UC), coverage probability (CP), confidence interval (CI) and average interval length (Length) of $\hat{\beta}_2$ and $\hat{\beta}_3$ on bootstrap procedures for $n = 500$.

% HLP	Estimation Methods	$\hat{\beta}_2$ LC	CP	UC	(CI), Length	$\hat{\beta}_3$ LC	CP	UC	(CI), Length
0	MLE	0	100	0	(-1.2752 - -0.7731) 0.5021	0	100	0	(-0.2113 - 0.2114) 0.4227
	CB	0	100	0	(-1.4094 - -0.6885) 0.7208	0	100	0	(-0.3043 - 0.3030) 0.6073
	DLGBB	0	100	0	(-1.4094 - -0.6885) 0.7208	0	100	0	(-0.3043 - 0.3030) 0.6073
	WLGBP	0	100	0	(-1.4094 - -0.6885) 0.7208	0	100	0	(-0.3043 - 0.3030) 0.6073
5	MLE	100	0	0	(-0.0846 - 0.1712) 0.2557	0	100	0	(-0.1803 - 0.1798) 0.3601
	CB	100	0	0	(-0.1449 - 0.2285) 0.3735	0	100	0	(-0.2593 - 0.2573) 0.5166
	DLGBB	0	100	0	(-1.4288 - -0.6735) 0.7553	0	100	0	(-0.3203 - 0.3213) 0.6416
	WLGBP	0	100	0	(-1.9221 - -0.9356) 0.9864	0	100	0	(-0.3854 - 0.3840) 0.7694
10	MLE	100	0	0	(-0.0406 - 0.2212) 0.2618	0	100	0	(-0.1845 - 0.1841) 0.3686
	CB	100	0	0	(-0.0971 - 0.2786) 0.3757	0	100	0	(-0.2641 - 0.2635) 0.5276
	DLGBB	0	100	0	(-1.4417 - -0.6682) 0.7734	0	100	0	(-0.3275 - 0.3306) 0.6581
	WLGBP	0	100	0	(-1.9384 - -0.9251) 1.0133	0	100	0	(-0.3959 - 0.3933) 0.7892
15	MLE	100	0	0	(-0.0200 - 0.2502) 0.2703	0	100	0	(-0.1891 - 0.1895) 0.3787
	CB	100	0	0	(-0.0769 - 0.3101) 0.3870	0	100	0	(-0.2722 - 0.2699) 0.5421
	DLGBB	0	100	0	(-1.4567 - -0.6614) 0.7953	0	100	0	(-0.3345 - 0.3358) 0.6703
	WLGBP	0	100	0	(-1.9555 - -0.9220) 1.0335	0	100	0	(-0.4012 -0.4010) 0.8022
20	MLE	100	0	0	(-0.0070 - 0.2712) 0.2782	0	100	0	(-0.1953 - 0.1960) 0.3913
	CB	100	0	0	(-0.0638 - 0.3343) 0.3981	0	100	0	(-0.2801 - 0.2812) 0.5613
	DLGBB	0	100	0	(-1.4723 - -0.6503) 0.8220	0	100	0	(-0.3446 - 0.3461) 0.6906
	WLGBP	0	100	0	(-1.9654 - -0.9122) 1.0532	0	100	0	(-0.4114 - 0.4081) 0.8195

Table 12. Estimated coefficients, standard errors, and the goodness of fit for the modified coronary heart disease data.

		Estimation Methods				
		MLE_100	MLE_97	CB	DLGBB	WLGBP
Intercept	Value	-1.4834	-5.1882	-2.0039	-5.3996	-7.3367
	Std. Err	0.6809	1.1540	1.3229	1.2925	1.5163
Age	Value	0.0254	0.1085	0.0370	0.1132	0.1535
	Std. Err	0.0138	0.0244	0.0290	0.0272	0.0317
	χ^2_{arc}	233.1438	182.4641	229.9521	181.5726	206.1502

Table 13. Estimated coefficients, standard errors, and the goodness of fit for the prostate cancer data.

		Estimation Methods				
		MLE_53	MLE_50	CB	DLGBB	WLGBP
Intercept	Value	1.3281	-0.1352	0.8890	-0.2502	-1.1275
	Std. Err	2.9030	3.3363	3.8591	3.9766	3.8480
AP	Value	0.0214	0.0414	0.0297	0.0450	0.0640
	Std. Err	0.0125	0.0200	0.0236	0.0247	0.0256
Age	Value	-0.0564	-0.0541	-0.0589	-0.0568	-0.0657
	Std. Err	0.0483	0.0511	0.0575	0.0580	0.0617
	χ^2_{arc}	114.4234	102.1756	112.9089	101.6271	110.1880

Table 14.Estimated coefficients, standard errors, and the goodness of fit for the intensive care unit data.

		Estimation Methods				
		MLE_200	MLE_198	CB	DLGBB	WLGBP
Intercept	Value	-1.0613	-0.7550	-1.1137	-0.8019	-2.1044
	Std. Err	1.2282	1.2750	1.2954	1.3623	1.7504
Age	Value	0.0284	0.0301	0.0303	0.0318	0.0580
	Std. Err	0.0108	0.0112	0.0112	0.0126	0.0181
SYS	Value	-0.0168	-0.0214	-0.0179	-0.0226	-0.0351
	Std. Err	0.0059	0.0063	0.0075	0.0063	0.0088
HRA	Value	0.0009	0.0021	0.0012	0.0025	0.0081
	Std. Err	0.0067	0.0068	0.0070	0.0070	0.0079
	χ^2_{arc}	300.443	286.6569	297.583	284.2126	277.2474

5/1/2021