Research Article GU J Sci 32(1): 318-331 (2019)

Gazi University

Journal of Science

http://dergipark.gov.tr/gujs

Comparison of Binary Logistic Regression Models Based on Bootstrap Method: An Application on Coronary Artery Disease Data

Hayriye Esra AKYUZ[1,*], Hamza GAMGAM[2]

[1] *Bitlis Eren University, Department of Statistics, 13000, Bitlis, Turkey*

[2] *Gazi University, Department of Statistics, 06500, Ankara, Turkey*

Article Info

Received: 22/04/2018
Accepted: 18/10/2018

Keywords

Bootstrap
Backward Elimination
Coronary Artery Disease
Error Term
Logistic Regression
Parameter Estimation

Abstract

This study is aimed to obtain an appropriate logistic regression model based on the bootstrap methods. For this purpose, two bootstrap methods called bootstrap I and bootstrap II are given to obtain the estimations of parameters and standard errors. Traditional logistic regression is compared with the bootstrap I and bootstrap II methods in terms of the parameter estimations and standard errors. It has been found that the standard errors of the parameter estimations for the bootstrap I model are smaller than others. Also, the average widths of confidence interval based on bootstrap I model are narrower than the logistic regression and bootstrap II. It is seen that, the simulation study based on different sample sizes supports these results. It can be said that the bootstrap I model based on resampling of errors term is the best in estimating coronary artery disease.

1. INTRODUCTION

The logistic regression (LR) analyzes the effect of the independent variable on the dependent variable by comparing the probability of occurrence of one of the two categories of dependent variable with the probability of occurrence of the other category. The purpose of LR analysis is to determine the model that explains the relationship between the least independent and dependent variable.

The researchers in the field of medicine also want to examine the effect of the factors on the dependent variable and their effects together. Another issue that is most important for the researchers is to examine the relationship between the factors and the disease in terms of risk. LR analysis is often used in such studies [1].

Binary Logistic Regression Analysis is based on probability ratios. The probability ratio compares the likelihood that an event will occur and the probability that it will not occur. The value of this regression analysis is obtained by taking the natural logarithm of the probability ratio. When model parameters are estimated, the maximum likelihood and Wald statistics are widely used [1-2].

Regression methods are commonly used to examine relationships between a dependent variable and one or more independent variables. The best known methods include simple and multiple linear regression, i.e., least squares methods, where the dependent variable is numerical. However, the dependent variable is often categorical for many studies, where least squares methods are unsuitable to obtain parameter estimates. In such cases, LR can be employed, and there are several LR models depending on the number of categories of dependent variables and whether the categories are nominal or ordinal [1]. LR analysis compares the

*Corresponding author, e-mail: heakyuz@beu.edu.tr

effects of independent variables on the dependent variable to likelihood of implementation of either of two categories [2].

Berkson [3] first proposed logistic models to analyse biological experiments. Lim et al. [4] were derived an additive model from transformation of the logistic model. Vupa and Çelikoğlu [5] proposed an LR model for lung cancer patients, and Coşkun et al. [6] applied LR in dentistry. LR has also been used in various previous studies to predict coronary artery disease (CAD) [7-9], and Atabey [10] applied LR for hospital patients with hypertension. Yan et al. [11] were studied the prediction of CAD in patients undergoing operations for rheumatic aortic valve disease. However, no previous study has estimated CAD using LR based on bootstrap sampling that applied both the error terms and independent variables.

CAD is a complex multifactorial disease characterized by various genes, environmental factors and interactions, and is the most common cardiac disease [12-14]. CAD is responsible for approximately one-third of all deaths in individuals 35 years and older [15]. The Turkish Adult Risk Factor Study Survey 2012 analyzed overall mortality and coronary mortality of the specific age group identified in follow-up screening of heart disease and risk factors [16]. When all causal mortalities for 45–74 aged patients in Turkey were considered, mortality was 16.8 per thousand for males, and 9.9 per thousand for females. Median mortality in the 45–74 age group in thirty European countries was 13.2 per thousand for males, and 7.3 per thousand for females. It is known that such rates are 30% in Turkey [16-18]. Therefore, this study employed LR based on bootstrap method to determine factors affecting CAD.

Bootstrap sampling method enables to estimate the standard error of a statistic for statistical inferences. In some cases, regression analysis assumptions may not be known, and bootstrap methods can be employed. In this study, it is aimed to obtain an appropriate model based on the LR analysis for the estimation of CAD. In addition to, the parameter estimates obtained in this model are compared with the parameter estimates based on the bootstrap method.

2. STATISTICAL ANALYSIS

This section explains how the LR and bootstrap methods were applied for this analysis.

2.1. Logistic Regression Analysis

LR is not concerned with estimating the dependent or response variable value. Rather, LR estimates the likelihood that the dependent variable is 1 (when the risk is determined as 1), which is between 0 and 1. When there are more independent variables in the model, the multiple LR method is used. Suppose the number of independent variables is *k*, and the vector of independent variables is $X=(X_1, X_2,..., X_k)$. Then the multiple LR model can be expressed as [1]:

$$\pi(x)=P(Y=1/x)=\frac{\exp(\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_kx_k)}{1+\exp(\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_kx_k)}, \quad (1)$$

$$=(1+\exp(-(\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_kx_k)))^{-1}, \quad (2)$$

where *Y* is the vector of binary response variable with Bernoulli distribution and β_i, $i=(0,1,2,...,k)$ are regression coefficients.

Equation (2) can be expressed as the odds of the dependent variable,

$$\frac{\pi(x)}{1-\pi(x)}=\exp(\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_kx_k), \quad (3)$$

where $\pi/(1-\pi)$ is described as odds ratio. We define the logit transformation is as

$$\text{logit}\pi(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right), \tag{4}$$

and taking the natural logarithm of Equation (3) provides a linear model,

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \tag{5}$$

where logit transformation is a linear function of β parameters. This is one of the assumption of the LR. Other assumptions of LR analysis are as follows [19]:

i. The response variable must be at least two levels.
ii. If the situation interested for response variable is occurred, it must be coded as "1".
iii. There should be no missing or excessive independent variables in the model.
iv. Each observation must be independent of each other.
v. There should not be a multicollinearity problem between the independent variables.
vi. To obtain strongest maximum likelihood estimators, the sample size should be at least 10 times the number of independent variables.

Multiple LR employs forward and backward elimination steps to determine the best fit model with the least number of variables. Forward selection starts with a model that includes only a constant, with no independent variables available. Then individual independent variables are progressively added, and the greatest change in log likelihood determined, to identify the best variable to include. The single parameter model is then similarly treated, and variable addition continues until the contributions of any remaining variables are insignificant. In contrast, backward elimination commences with all variables included, and individual variables are removed, to identify the variable that caused the least increase in deviation, which is then removed from the model. This process continues until the variable that caused a significant change in deviation is obtained when removed from the model [1-2].

2.2. Bootstrap Method

The bootstrap sample method estimates the standard error of a statistic and obtains confidence intervals and distributions. The method assumes that the observed data is representative of the population under examination [20].

Most statistical problems are concerned with the distribution of a statistic of a random sample drawn from the population. If the distribution is known, it is possible to apply general theories to generate the sample distribution. When there is no significant information about other features of the distribution or the estimator, the bootstrap method is used.

In the bootstrap method, the original sample is substituted by a bootstrap sample and resampled. The sample is treated like a real population, and this sampling is repeated many times to create an experimental distribution for the estimator [21]. Population parameters are estimated by minimizing the estimated standard error from the sample data [22].

In the bootstrap method, sample of size n are selected from the data set $x = (x_1, x_2, ..., x_n)$ for the standard error of estimator $\hat{\theta}$, and *B* bootstrap samples are created, $x_1^*, x_2^*, ..., x_B^*$. The asterisks in the bootstrap sample is created by substituting it in the observed real values in the sampling method. The bootstrap

estimate, $\hat{\theta}^*$, is obtained for θ, and $x_1^*, x_2^*, ..., x_B^*$ is repeated *B* times in obtaining a bootstrap sample and bootstrap estimate $\hat{\theta}^*$. Estimates $\hat{\theta}^{*1}, \hat{\theta}^{*2}, ..., \hat{\theta}^{*B}$ are obtained after the *B* replications, and

$$s_{\hat{\theta},boot}^2 = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}^{*b} - \bar{\hat{\theta}}^*\right)^2 \tag{6}$$

where $\bar{\hat{\theta}}^*$ is the average of $\hat{\theta}^*$,

$$\bar{\hat{\theta}}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*b} \tag{7}$$

and the square root of the $s_{\hat{\theta},boot}^2$ is defined as the standard error of the estimator for the bootstrap samples [23].

2.3. Logistic Regression Analysis Based on the Bootstrap Method

Assume a random sample of size *n* as $(x_1, x_2, ..., x_n)$. Let $x_i = (c_i, y_i)$ for each x_i observation, where y_i is the dependent variable and c_i is the $1 \times k$ vector formed by the independent variables, i.e., $c_i = c_{i1}, c_{i2}, ..., c_{ik}$. If the values of independent variable are known, then the expected value of the dependent variable is obtained as:

$$\mu_i = E(y_i / c_i),\ i = 1, 2, ..., n \tag{8}$$

where y_i is the dependent variable and μ_i is a linear function of c_i independent variables. Also, μ_i is as follows:

$$\mu_i = c_i\beta = \sum_{j=1}^{p} c_{ij}\beta_j\ . \tag{9}$$

The purpose of the analysis is to estimate unknown parameters $\beta = (\beta_1, \beta_2, ..., \beta_k)$ using the observation data $(x_1, x_2, ..., x_n)$. Linear model is as follows:

$$y_i = c_i\beta + \varepsilon_i,\quad i = 1, 2, ..., n\,. \tag{10}$$

When the expected value of the error terms ε_i is zero and they have an unknown *F* distribution, it is expressed as [24]:

$$F \to (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n) = \varepsilon,\quad E(\varepsilon) = 0 \tag{11}$$

From Equation (10) and Equation (11), we have,

$$E\left(y_i / c_i\right) = E\left(c_i\beta + \varepsilon_i / c_i\right) = E\left(c_i\beta / c_i\right) + E\left(\varepsilon_i / c_i\right)$$
$$= c_i\beta$$

This expression is the linearity assumption in Equation (10). Parameter estimates in Equation (10) is obtained with least squares method and they are as follows:

$$\hat{\beta} = \left(C^T C\right)^{-1} C^T y\,. \tag{12}$$

The bootstrap method can be used in two different ways for regression analysis. These methods are described as follows.

The probability model of the linear regression is P → x where $P = (\beta, F)$, F is the distribution of the error terms, and β is the regression coefficients [24]. Since β is unknown, estimates $\hat{\beta}$ obtained using least squares method are used and experimental distributions of error terms obtained. To obtain a bootstrap estimator of β, the algorithm is used as follows:

- Random sample from population is selected.
- The regression model based on this sample is obtained.
- $\hat{\varepsilon}_i = y_i - c_i\hat{\beta}, \quad i = 1, 2, ..., n$ are calculated.
- $\hat{\varepsilon}_i$ values are given 1 / n probability to obtain bootstrap error sub-sample B, each in size n. Thus, emprical distribution function of error terms $\left(\hat{F}_\varepsilon\left(x\right)\right)$ is as:

  $\hat{F}_\varepsilon\left(x\right) = \left\{\hat{\varepsilon}_i^{\,2} \le x\right\} \Big/ n$
- Average of bootstrap error values is $\bar{\hat{\varepsilon}}_i^{\,*} = \sum_{b=1}^{B} \hat{\varepsilon}_{bi} \Big/ B$.
- Estimates of dependent variables are $\hat{y}_i = c_i\hat{\beta} + \hat{\varepsilon}_i$. If $\bar{\hat{\varepsilon}}_i^{\,*}$ is substituted in this equation, the bootstrap value Y_i^* are as follows:

  $Y_i^* = C\hat{\beta} + \bar{\hat{\varepsilon}}_i^{\,*}$
- Bootstrap estimator $\hat{\beta}^*$ is as:

$$\hat{\beta}^* = \left(C^T C\right)^{-1} C^T y^* \tag{13}$$

where C is matrix of independent variables. As a result, it is used bootstrap approach based on resampling of error terms. We call this method as bootstrap I.

On the other hand, the bootstrap method can be applied to the data set $x_i = (c_i, y_i)$. In this case, bootstrap data set is as follows:

$$x^* = \left\{(c_{i1}, y_{i1}), (c_{i2}, y_{i2}), ..., (c_{in}, y_{in})\right\} \tag{14}$$

where c_i are independent variables, y is dependent variable. The regression coefficients for each bootstrap sample can be estimated as:

$$\hat{\beta}^* = \left(C^{*T} C^*\right)^{-1} C^{*T} y^*. \tag{15}$$

The expected value of the bootstrap estimators is approximately equal to the least squares estimator of β [25].

Similarly, we call this method as bootstrap II. While the bootstrap I method works by resampling the error terms, bootstrap II method works by resampling of the observation values (c_i, y_i) where dependent variable y_i and vector formed by the independent variables c_i.

3. MATERIAL AND METHODS

Appropriate models were obtained using backward elimination, with least squares for parameter estimates, and experimental distributions of the error terms were determined from the $\hat{\beta}$ estimates. The data were officially obtained from General Union of Bitlis Public Hospitals for 170 persons, including 85 healthy and 85 CAD patients, from the Cardiology Department of Bitlis State Hospital between 2012 and 2017. For this study, necessary permissions were obtained from the ethics committee of the Bitlis Eren University. LR was applied to the data to provide a statistical model the patient's CAD risk, using statistical softwares R and SPSS . Independent variables in the model was determined by SPSS package program and the bootstrap method was implemented with the R package program. Independent variables used to estimate CAD included age, hemoglobin, white blood cell (WBC), uric acid, high density lipoprotein (HDL), low density lipoprotein (LDL), triglyceride, total bilirubin, and direct bilirubin. The variables that make significant contribution to the model were determined by backward elimination for the CAD dependent variable. Parameter estimates were obtained by applying the bootstrap technique for *B*=500, 1000, 1500 and 2000 replications on the same dataset. Finally, estimated parameter values obtained from LR were compared to the estimated parameters values , obtained from the bootstrap method.

On the other hand; a simulation study is conducted and is compared with the results based on CAD. The dependent variable consists of two categories and independent variables have normal distribution with mean 0 and variance 1. These are obtained by deriving 1000 replications data when all of the independent variables are continuous and sample size n=20, 50, 100, 500.

Using all the observations of the population will cause a time waste and increase the cost. The bootstrap method eliminates these problems and offers great advantages. This study investigated if CAD risk obtained from the bootstrap method achieved more efficiency parameter estimates compared to those obtained from LR, compared to standard errors. In this study, $Y_i = 1$ means the patient has CAD, and $Y_i = 0$ means complete absence of CAD.

4. RESULTS

Table 1. *Some descriptive statistics for the independent variables*

Variables	Mean ±Standard Deviation	Min	Max
Age	55.74±8.33	38	78
Hemoglobin(mg/dl)	13.73±2.02	9	25.50
WBC (mg/dl)	7363.35±1506.69	4600	14000
Uric acid (mg/dl)	5.01±1.32	2.60	9
HDL (mg/dl)	37.23±8.08	19	58
LDL (mg/dl)	127.47±25.06	61	202
Triglyceride (mg/dl)	138.64±47.52	41	274
Total bilirubin (mg/dl)	0.76±0.27	0.10	1.30
Direct bilirubin (mg/dl)	0.17±0.09	0.01	0.50

Table 1 shows some descriptive statistics of independent variables used in CAD estimation. The age range of the 170 patients used in the current study varied between 38–78, with average age = 55. Average hemoglobin = 13.73 ± 2.02, WBC = 7363.35 ± 1506.69; uric acid = 5.01 ± 1.32; HDL = 37.23 ± 8.08; LDL

= 127.47 ± 25.06; triglyceride = 138.64 ± 47.52; total bilirubin = 0.76 ± 0.27; and direct bilirubin = 0.17 ± 0.09.

The regression procedures for categorical dependent variables do not have multicollinearity diagnostics. However, we can use the linear regression procedure for this purpose. Multicollinearity is particularly problematic for logistic regression models. It usually occurs when one or more independent variables are related to each other.

Table 2. *Correlations between independent variables*

	Age	Hemoglobin	WBC	Uric acid	HDL	LDL	Triglyceride	Total bilirubin	Direct bilirubin
Age	1	0.009	0.118	0.054	0.061	0.242**	0.185*	0.051	0.057
		0.912	0.126	0.484	0.429	0.001	0.016	0.512	0.462
Hemoglobin	0.009	1	0.091	0.194*	-0.025	0.040	0.162*	0.183*	0.089
	0.912		0.240	0.011	0.747	0.601	0.035	0.017	0.246
WBC	0.118	0.091	1	0.059	-0.006	0.209**	0.418**	0.074	0.254**
	0.126	0.240		0.447	0.937	0.006	0.000	0.340	0.001
Uric acid	0.054	0.194*	0.059	1	0.191*	0.060	0.299**	0.084	0.067
	0.484	0.011	0.447		0.012	0.439	0.000	0.276	0.384
HDL	0.061	-0.025	-0.006	0.191*	1	0.060	0.052	0.182*	0.104
	0.429	0.747	0.937	0.012		0.436	0.503	0.017	0.176
LDL	0.242**	0.040	0.209**	0.060	0.060	1	0.400**	0.162*	-0.009
	0.001	0.601	0.006	0.439	0.436		0.000	0.035	0.909
Triglyceride	0.185*	0.162*	0.418**	0.299**	0.052	0.400**	1	0.389**	0.223**
	0.016	0.035	0.000	0.000	0.503	0.000		0.000	0.003
Total bilirubin	0.051	0.183*	0.074	0.084	0.182*	0.162*	0.389**	1	0.534**
	0.512	0.017	0.340	.276	.017	.035	.000		.000
Direct bilirubin	0.057	0.089	0.254**	0.067	0.104	-0.009	0.223**	0.534**	1
	0.462	0.246	0.001	0.384	0.176	0.909	0.003	0.000	

**: Correlation is significant at the 0.01 level, *: correlation is significant at the 0.05 level

High correlation between the independent variables is an undesirable. It is a potential multicollinearity condition. Correlations between independent variables are obtained in Table 2. When the correlation matrix for the independent variables is examined, it is seen that there are no highly correlated variables.

Another way to diagnose the multipcollinearity presence is to run logistic regression as a linear regression by putting one of the independent variables in the model as dependent variable. In Table 3, the tolerance value, variance inflation factor (VIF), condition index and eigenvalue is checked for multipcollinearity diagnostics.

Table 3. *Values of tolerance, variance inflation factor (VIF), condition index and eigenvalue*

Dependent variable	Independent variable	Criteria of multipcollinearity			
		Tolerance	VIF	Condition index	Eigenvalue
Age	Hemoglobin	0.920	1.087	6.581	0.197
	WBC	0.741	1.349	10.116	0.183
	Uric acid	0.836	1.196	11.430	0.165
	HDL	0.919	1.088	12.452	0.155
	LDL	0.813	1.230	15.584	0.135
	Triglyceride	0.562	1.779	17.861	0.127
	Total bilirubin	0.569	1.756	20.661	0.120
	Direct bilirubin	0.644	1.552	34.921	0.107
Hemoglobin	Age	0.926	1.080	33.526	0.108
	WBC	0.744	1.344	6.554	0.198
	Uric acid	0.864	1.157	10.104	0.183
	HDL	0.925	1.082	11.436	0.165
	LDL	0.784	1.275	12.213	0.157
	Triglyceride	0.559	1.790	16.272	0.132
	Total bilirubin	0.582	1.719	18.082	0.126
	Direct bilirubin	0.644	1.554	20.511	0.120

Table 3 continued

WBC	Age	0.926	1.080	21.279	0.119
	Hemoglobin	0.923	1.083	37.035	0.106
	Uric acid	0.843	1.186	6.515	0.201
	HDL	0.917	1.091	10.003	0.185
	LDL	0.789	1.267	11.782	0.161
	Triglyceride	0.653	1.531	12.987	0.151
	Total bilirubin	0.604	1.657	15.521	0.135
	Direct bilirubin	0.699	1.430	18.986	0.124
Uric acid	Age	0.925	1.081	19.200	0.123
	Hemoglobin	0.951	1.052	21.283	0.119
	WBC	0.747	1.338	38.279	0.106
	HDL	0.959	1.043	6.602	0.196
	LDL	0.787	1.270	10.029	0.185
	Triglyceride	0.614	1.629	11.462	0.165
	Total bilirubin	0.575	1.739	15.272	0.137
	Direct bilirubin	0.644	1.553	17.508	0.128
HDL	Age	0.927	1.078	17.604	0.128
	Hemoglobin	0.927	1.078	20.213	0.121
	WBC	0.741	1.350	21.283	0.119
	Uric acid	0.874	1.144	37.088	0.106
	LDL	0.786	1.273	6.548	0.199
	Triglyceride	0.563	1.777	10.346	0.180
	Total bilirubin	0.584	1.713	11.759	0.162
	Direct bilirubin	0.643	1.556	12.267	0.157
LDL	Age	0.959	1.042	12.675	0.153
	Hemoglobin	0.920	1.087	15.525	0.135
	WBC	0.746	1.341	19.209	0.123
	Uric acid	0.840	1.191	21.232	0.119
	HDL	0.919	1.088	37.693	0.106
	Triglyceride	0.612	1.635	6.615	0.195
	Total bilirubin	0.572	1.747	10.027	0.185
	Direct bilirubin	0.659	1.518	11.432	0.165
Triglyceride	Age	0.931	1.074	12.211	0.157
	Hemoglobin	0.920	1.087	14.450	0.141
	WBC	0.867	1.154	17.663	0.127
	Uric acid	0.920	1.087	18.045	0.126
	HDL	0.924	1.082	21.432	0.119
	LDL	0.859	1.164	37.471	0.106
	Total bilirubin	0.644	1.554	6.508	0.202
	Direct bilirubin	0.645	1.550	11.274	0.167
Total bilirubin	Age	0.928	1.078	10.176	0.183
	Hemoglobin	0.942	1.062	12.241	0.157
	WBC	0.787	1.270	15.112	0.137
	Uric acid	0.846	1.182	17.226	0.129
	HDL	0.942	1.061	19.378	0.123
	LDL	0.790	1.266	21.046	0.119
	Triglyceride	0.632	1.581	38.470	0.106
	Direct bilirubin	0.891	1.122	6.805	0.185
Direct bilirubin	Age	0.928	1.078	9.274	0.101
	Hemoglobin	0.921	1.086	10.707	0.175
	WBC	0.806	1.241	12.302	0.157
	Uric acid	0.837	1.194	15.572	0.136
	HDL	0.917	1.091	16.813	0.131
	LDL	0.803	1.245	20.229	0.121
	Triglyceride	0.560	1.785	21.312	0.119
	Total bilirubin	0.787	1.270	38.716	0.106

The VIF values above 5 or 10 are indicative of a strong multicollinearity. The fact that the condition index value is below 100 indicates that there is no serious multicollinearity problem in the data [26]. If there is no multicollinearity, the correlation coefficient between the independent variables will be low and tolerance value approaches to 1. When the results in Table 3 are examined, it is seen that there is no multicollinearity problem in the data set.

Table 4. *Determination of independent variables related to coronary artery disease using backward elimination*

Step	Variables	Coefficients	S.E.	Wald	df	*p*-values	Odds ratio	95% Confidence Interval	
								Lower	Upper
7	Age	0.154	0.040	15.245	1	<0.001	1.167	1.080	1.261
	LDL	0.070	0.017	17.934	1	<0.001	1.073	1.039	1.108
	Triglyceride	0.054	0.010	26.825	1	<0.001	1.055	1.034	1.077
	Constant	-25.202	4.143	37.004	1	<0.001	0.000		

*: S.E.:standard error, df: degree of freedoom.

For the multiple LR model, the significance level, α, was used to select the variables to be included in the model, in contrast to linear regression. Selection of the *p*-value has been demonstrated by Bendel and Afifi [27] for linear regression and Mickey and Greenland [28] for LR. This study chose $\alpha = 0.25$, and backward elimination based on the likelihood ratio test statistic was used to establish the best model.

Table 4 shows the independent variables included in the model. All variables are included in the model in step 1. By using of the backward elimination method, HDL in step 2 and uric acid in step 3 are eliminated from model. This process continues until step 7, where no variables are identified as insignificant, i.e., removed from the model. The remaining variables: age, LDL, and triglyceride, are identified as the significant predictors for the model.

Table 5. *Hosmer-Lemeshow and model coefficient tests*

	Chi-square	df*	*p*-value
Hosmer-Lemeshow test	5.721	8	0.678
Model coefficient test	156.554	9	<0.001

*: degree of freedom

When Table 5 is examined, it is determined that the Chi-square values are statistically significant (*p*-value<0.05). The significant Chi-square statistic implies rejection of the null hypothesis: there is no difference between the initial model (including just the constant term) and the final model (including the identified independent variables). Thus, the LR coefficients except for the constant are different from zero, i.e., all model coefficients are significant and there is a significant relationship between the dependent variable and these independent variables. In addition to; Table 5 shows that the theoretical model represents the data well (*p*-values> 0.05). This test evaluates compatibility of model as a whole.

Table 6. *Classification table*

			Predicted		
Step	Observed		Healthy	Disease	Percentage Correct
7	Coronary Artery Disease	Healthy (0)	77	8	90.6
		Disease (1)	7	78	91.8
	Overall Percentage				91.2

Table 6 shows that 90.6% of healthy persons and 91.8% of CAD patients were correctly estimated in Step 7. In general, 91.2% were correctly estimated, which is very acceptable.

Table 7. *Model summary*

Step	-2 Log Likelihood	Cox & Snell R Square	Nagelkerke R Square
7	83.900	0.590	0.787

Table 7 shows that the independent variables explain 59.0% and 78.7% of total variation in the dependent variable according to Cox-Snell and Nagelkerke, respectively, in Step 7, and the CAD contributions for independent variables in subsequent steps.

Table 8. *Prediction equation variables*

Variable	Coefficient	Standard Error	Wald	df	*p*-value	Odds ratio	95% Confidence Interval	
							Lower	Upper
Age	0.154	0.040	15.245	1	<0.001	1.167	1.080	1.261
LDL	0.070	0.017	17.934	1	<0.001	1.073	1.039	1.108
Triglyceride	0.054	0.010	26.825	1	<0.001	1.055	1.034	1.077
Constant	-25.200	4.143	37.004	1	<0.001	0.000		

In Table 8, according to odds ratios; CAD incidence increases 1.167, 1.073, and 1.055 times as age, LDL and triglyceride increases, respectively. Thus age, LDL and triglycerides are the important independent variables determining CAD, and the LR model was obtained as:

$$P(CAD=1)=(1+e^{-(-25.200+0.154age+0.070LDL+0.054triglyceride)})^{-1} \quad . \tag{13}$$

Table 9. *Parameter estimations and standard errors based on LR and bootstrap methods for different replication numbers*

Original Logistic Regression			Bootstrap I		Bootstrap II	
Variable	Estimation	Standard Error	Estimation	Standard Error	Estimation	Standard Error
			B=500*			
Age	0.154	0.040	0.152	0.031	1.347	0.130
LDL	0.070	0.017	0.070	0.009	0.310	0.025
Triglyceride	0.054	0.010	0.049	0.009	0.039	0.020
Constant	-25.200	4.143	-24.868	3.880	-27.656	4.578
			B=1000*			
Age	0.154	0.040	0.152	0.031	1.312	0.125
LDL	0.070	0.017	0.069	0.009	0.309	0.021
Triglyceride	0.054	0.010	0.049	0.009	0.031	0.019
Constant	-25.200	4.143	-24.937	3.964	-27.678	4.569
			B=1500*			
Age	0.154	0.040	0.152	0.031	1.310	0.120
LDL	0.070	0.017	0.070	0.009	0.310	0.020
Triglyceride	0.054	0.010	0.050	0.009	0.030	0.020
Constant	-25.200	4.143	-24.930	3.960	-27.661	4.560
			B=2000*			
Age	0.154	0.040	0.152	0.031	1.310	0.120
LDL	0.070	0.017	0.070	0.009	0.310	0.020
Triglyceride	0.054	0.010	0.050	0.009	0.030	0.020
Constant	-25.200	4.143	-24.930	3.960	-27.661	4.560

*: B is bootstrap replication number

Table 9 shows the parameter estimations and standard errors for the *B*= 500, 1000, 1500 and 2000 replications bootstrap sampling and backward elimination method. In Table 9, it is seen that the parameter estimations of the bootstrap I are nearly close to that of the LR method, whereas those from the bootstrap II method are larger. In addition to, the standard errors for the bootstrap I are lower than those of the LR method, althought the standard errors for the bootstrap II are larger than those of others. As expected, when the number of bootstrap replications increased, it is obtained that the standard error values of the bootstrap methods decreased. The LR model based on the bootstrap I samples with *B*=2000 replications is as follows:

$$P(CAD=1)=(1+e^{-(-24.930+0.152age+0.070LDL+0.050triglyceride)})^{-1} . \tag{14}$$

It is seen that the coefficients of this model are quite similar to the coefficients of the original LR model. As a result; the standard error values of estimations decreased and more efficiency parameter estimations are obtained.

Table 10*. Confidence intervals of parameters and odds ratios for the LR and bootstrap methods*

	LR	Bootstrap I	Bootstrap II
Parameter	OR (95% CI)	OR (95% CI)	OR (95% CI)
β_{age}	1.167 (1.080-1.261)	1.367 (1.083-1.210)	1.379 (1.362-1.525)
β_{LDL}	1.073 (1.039-1.108)	1.098 (1.046-1.100)	1.284 (1.221-1.290)
$\beta_{Triglyceride}$	1.055 (1.034-1.077)	2.169 (2.386-2.425)	2.287 (2.258-2.300)

OR: odds ratio, CI: confidence interval

In Table 10, LR and bootstrap methods were compared in terms of confidence intervals for parameters and odds ratios. The odds ratios obtained by both bootstrap methods are quite similar to the LR method. But, it is easily seen that the widths of the confidence interval obtained by the bootstrap I method are narrower then others.

Table 11. *Parameter estimations and standard errors for data generated normal distribution N(0,1)*

n	Variable	Logistic Regression		Bootstrap I		Bootstrap II	
		Estimation	Standard Error	Estimation	Standard Error	Estimation	Standard Error
20	X_1	0.040	0.150	0.049	0.114	0.098	0.265
	X_2	0.089	0.155	0.092	0.127	0.125	0.196
	X_3	0.053	0.159	0.040	0.114	0.127	0.241
	Constant	0.003	0.154	0.001	0.101	0.095	0.189
50	X_1	0.056	0.125	0.044	0.111	0.365	0.142
	X_2	1.092	0.149	1.009	0.138	1.896	0.186
	X_3	1.065	0.160	1.096	0.143	1.678	0.186
	Constant	0.001	0.142	0.001	0.109	0.075	0.175
100	X_1	0.156	0.090	0.156	0.089	1.954	0.152
	X_2	1.198	0.083	1.197	0.080	1.969	0.154
	X_3	0.167	0.040	0.167	0.030	1.786	0.102
	Constant	0.010	0.024	0.010	0.024	0.147	0.147
500	X_1	0.193	0.025	0.195	0.022	1.896	0.025
	X_2	1.945	1.009	1.947	0.989	1.996	0.999
	X_3	1.563	0.965	1.565	0.897	1.945	0.899
	Constant	0.047	0.027	0.047	0.019	0.563	0.020

In Table 11, a simulation study with 1000 replications is conducted with two categories dependent variable and independet variables generated standard normal distribution as in the study of Stute et. al. [29] and Karadağ [30]. On the other hand; as the LR model obtained by the backward elemination method has three independent variables, this independent variables (X_1, X_2, X_3) are generated in the simulation study. According to Table 11, parameter estimates and standard errors based on both bootstrap methods are quite similar with LR method for all sample size. When sample sizes is n=20, 50, 100, standard errors of parameter estimations based on botstrap I are smaller than those of bootstrap II. However, the standard error values based on both methods are very close to each other for n=500. The simulation study support the results obtained with CAD data.

5. CONCLUSIONS

In LR, independent variables can be continuous or categorical, and there are no restrictive assumptions on independent variable distributions, e.g. normally distributed. These are major advantages for non-normally distributed datasets, and LR is becoming extensively employed in many medical studies. Therefore, this study developed an LR model for the presence of CAD in hospital patients.

LR identifies the best fit for the dependent variable using the least number of candidate independent variables. The effects are calculated as probabilities and in this case the final outcome is an overall CAD

risk or probability. This study determined that the significant CAD predicters were age, LDL and triglycerides. Thus, with fewer variables, time was earned, facilitating transaction. After this step, samples were obtained by the bootstrap method which is one of the resampling methods and it was tried to be shown which of the more efficiency parameter estimates were obtained by comparing with the model obtained by the LR analysis.

The study shows that the prevalence of CAD increases by 1.167, 1.073, and 1.055 times as age LDL and triglyceride increased, and the final prediction relationship was significant (p-value < 0.05).

For the bootstrap method in this study, new data sets of the same size (n=170) were generated on the data obtained from 170 persons by applying bootstrap resampling method. In the data sets, the number of the bootstrap replications was taken as B = 500, 1000, 1500 and 2000, and LR models were obtained with the parameter estimations.

Bootstrap methods measure the power to represent the population of the data set. When the results of the LR analysis and the bootstrap method are examined, it is seen that the parameter estimations of the bootstrap I are nearly close to that of the LR method, but the estimations for the bootstrap II are larger than those of the LR and bootstrap I. As a result of the evaluation of standard errors, it can be said that the standard errors for the bootstrap I are lower than those of the LR and bootstrap II, and the standard errors of the bootstrap II are the larger. On the other hand, bootstrap methods may not always produce effective results, the success of the method depends on the data structure and that the experimental distribution reflects the population distribution well.

When both bootstrap methods are examined odds ratios and confidence interval, it is seen that the odds ratios obtained by bootstrap methods are quite similar to the LR method. The widths of the confidence interval obtained by the bootstrap I method are narrower then others.

According to the results of the study, it is determined that the parameter estimates obtained from the dataset (n=170) in which bootstrap I method is applied are more effective than the parameter estimates of the original LR model. It is seen that the simulation study support the results based on CAD data.

CONFLICTS OF INTEREST

No conflict of interest was declared by the authors.

REFERENCES

[1] Alpar, R., Uygulamalı Çok Değişkenli İstatistiksel Yöntemler, Detay Yayıncılık, Ankara; (2011).

[2] Kleinbaum, D., Klein, M., Logistic Regression- A Self Learning Text, II ed. New York, NY: Springer; (2002).

[3] Berkson, J., "Application of the logistic function to bio-assay", J. Am. Stat. Assoc., 39(227): 357–65, (1944).

[4] Lim, E., Ali, Z.A., Barlow, C.W., Jackson, C.H., Hosseinpour, A.R., Halstead, J.C., et al., "A simple model to predict coronary disease in patients undergoing operation for mitral regurgitation", Ann. Thorac. Surg., 75(6):1820–5, (2003).

[5] Vupa, O., Çelikoğlu, C., "Model building in logistic regression models about lung cancer data", Anadolu Univ. J. Sci. Tech., 7(1): 127–41, (2006).

[6] Coşkun, S., Kartal, M., Coşkun, A., Bircan, H., "Lojistik regresyon analizinin incelenmesi ve diş hekimliğinde bir uygulaması", Cumhuriyet Üniversitesi Diş Hekimliği Fakültesi Dergisi, 7(1): 42–50, (2004).

[7] Hirashiki, A., Yamada, Y., Murase, Y., Hirashiki, A., Yamada, Y., Murase, Y., "Association of gene polymorphisms with coronary artery disease in low- or high-risk subjects defined by conventional risk factors", J. Am. Coll. Cardiol., 42(8): 1429–37, (2003).

[8] Horibe, H., Yamada, Y., Ichihara, S., Watarai, M., Yanase, M., Takemoto, K., et al., "Genetic risk for restenosis after coronary balloon angioplasty", Atherosclerosis, 174(1): 181–7, (2004).

[9] Çolak, C., Çolak, M.C., Orman, M.N., "The Comparison of logistic regression model selection methods for the prediction of coronary artery disease", The Anatol. J. Cardiol., 7(1): 6–12, (2007).

[10] Atabey, Ö. "Lojistik regresyon modeli ve geriye doğru eliminasyon yöntemiyle değişken seçiminin hipertansiyon riski üzerine uygulamasında bootstrap yöntemi", Gazi Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, Ankara, (2010).

[11] Yan, T., Zhang, G.X., Li, B.L., Han, L., Zang, J.J., Li, L., Xu, Z.Y., "Prediction of coronary artery disease in patients undergoing operations for rheumatic aortic valve disease", Clin. Cardiol., 35(11): 707-11, (2012) .

[12] Gündoğdu, F., Özdemir, Ö., Sevimli, S., Açıkel, M., Pirim, İ., Karakelleoğlu, Ş., et al., "The relationship between interleukin-6 polymorphism and the extent of coronary artery disease in patients with acute coronary syndrome", Arch. Turk. Soc. Cardiol., 35(5): 278–83, (2007).

[13] Yin, Y., Li, J., Zhang, M., Wang, J., Li, B., Liu, Y., et al., "Influence of interleukin-6 gene -174g>c polymorphism on development of atherosclerosis: a meta-analysis of 50 studies involving, 33.514 subjects", Gene, 529: 94–103, (2013).

[14] Elsaid, A., Abdel-Aziz, A.F., Elmougy, R., Elwaseef, A.M., "Association of polymorphisms G (-174) C in IL-6 gene and G (-1082) A in IL-10 gene with traditional cardiovascular risk factors in patients with coronary artery disease", Indian J. Biochem. Biophys., 51: 282–92, (2014).

[15] Roger, V.L., Go, A.S., Lloyd-Jones, D.M., Benjamin, E.J., Berry, J.D., Borden, W.B., et al., "Executive summary: heart disease and stroke statistics-2012 update: a report from the american heart association", Circ., 125(1): 188–97, (2012).

[16] Onat, A., Yüksel, M., Köroğlu, B., Gümrükçüoğlu, H.A., Aydın, M., Çakmak, H.A., "Turkish adult risk factor study survey 2012: overall and coronary mortality and trends in the prevalence of metabolic syndrome", Arch. Turk. Soc. Cardiol., 41: 373–8, (2013).

[17] Anderson, D.R., Poterucha, J.T., Mikuls, T.R., Duryee, M.J., Garvin, R.P., Klassen, L.W., et al., "IL-6 and its receptors in coronary artery disease and acute myocardial infarction", Cytokine, 62(3): 395-400, (2013).

[18] Teixeira, B.C., Lopes, A.L., Macedo, R.C.O., Correa, C.S., Ramis, T.R., Ribeiro, J.L., et al., "Inflammatory markers, endothelial function and cardiovascular risk", J. Vascul. Brasileiro, 13(2): 108–15, (2014).

[19] Hosmer, D., Lemeshow, S., Sturdıvant, R., Applied Logistic Regression. Canada: Wiley&Sons Publications, (2013).

[20] Mammen, E., When Does Bootstrap Work?. USA: Springer-Verlag New York, (1992).

[21] Özdemir, A., "Doğrusal olmayan regresyonda asimptotik yöntemle bootstrap örneklemesi", Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul, (2011).

[22] Aktükün, A. "Asal bileşenler analizinde bootstrap yaklaşımı", İstanbul Üniversitesi İktisat Fakültesi Ekonomi ve İstatistik Dergisi, 1: 1–11, (2005).

[23] Chernick, M.R., Bootstrap Methods. (2nd edition). Canada: John Wiley and Sons, (1999).

[24] Efron, B., Tibshirani RJ. An Introduction to the Bootstrap. Chapman & Hall, New York USA, (1993).

[25] Shao, J., Tu D, The Jacknife and Bootstrap, Spinger-Veriag, Newyork, (1995).

[26] Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E., Tatham, R.,L. Multivariate Data Analysis, Upper Saddle River, NJ: Pearson Prentice Hall, (2006).

[27] Bendel, R.B., Afifi, A.A., "Comparison of stopping rules in forward stepwise regression", J. Am. Stat. Assoc., 72(357): 46–53, (1977).

[28] Mickey, R.M., Greenland, S., "The impact of confounder selection criteria on effect estimation", Am. J. Epidemiol., 129(1): 125–37, (1989).

[29] Stute, W., Manteiga, W. G., Quindimil, M. P. "Bootstrap approximations in model checks for regression", Journal of the American Statistical Association, 93(441):141–9, (1998).

[30] Karadağ, M., "Karar ağaçları ile lojistik regresyon analizinin performanslarının simülasyon çalışması ile karşılaştırılması", Yüksek Lisans Tezi, Trakya Üniversitesi, Sağlık Bilimleri Enstitüsü, Edirne, (2014).