VERİ BİLİMİ DERGİSİ

www.dergipark.gov.tr/veri

LASSO Estimator in Logistic Regression for Small Data Sets

Aslı YAMAN[1]*, Mehmet Ali CENGİZ[1]

[1]*Ondokuz Mayıs Üniversitesi, Fen Edebiyat Fakültesi, İstatistik Bölümü, Samsun*

Abstract

Variable selection is an important subject in regression analysis. In regression analysis, the LASSO (Least Absolute Shrinkage and Selection Operator) provides sparse solutions to lead to variable selection. LASSO is a useful tool to achieve the shrinkage and variable selection simultaneously and the LASSO penalty term can shrink the parameter estimates toward exactly to zero. It is used generally in large data sets but in this article, we consider the variable selection problem for the multivariate Bernoulli logistic models adopting some information criteria especially in small data sets. Results of simulation were compared according to the four different criteria used for model selection.

Keywords: LASSO, Bernoulli distribution, logistic regression, feature selection

Makale Bilgisi

Başvuru:
22/10/2020
Kabul:
17/11/2020

Lojistik Regresyon Modellerinde Küçük Veri Setleri için LASSO Tahmincisi

Özet

Değişken seçimi, regresyon analizinde kullanılan önemli bir konudur. Regresyon analizinde, LASSO (En Küçük Mutlak Daralma ve Seçim Operatörü) değişken seçimine benzer olarak seyrek çözümler sunmaktadır. LASSO, daraltma ve değişken seçimi işlemlerini aynı anda yapabilen kullanışlı bir araçtır ve LASSO ceza kriteri, parametre tahminlerini tam olarak sıfır değerine indirebilir. Genellikle büyük veri kümelerinde kullanılır fakat bu çalışmada, özellikle küçük veri setlerinde bazı bilgi kriterlerini kullanarak çok değişkenli Bernoulli lojistik modelleri için değişken seçim problemi ele alınmıştır. Model seçiminde kullanılan dört farklı bilgi kriterine göre elde edilen simülasyon sonuçları karşılaştırılmıştır.

Anahtar Kelimeler: LASSO, Bernoulli dağılımı, lojistik regresyon, değişken seçimi

***** İletişim e-posta: asliyamann@gmail.com

1 Introduction

One of the statistical methods used in model selection is variable selection. Variable selection is a widely used method based on the Ordinary Least Squares (OLS) estimation method. However, this method is usually insufficient in the data sets where the number of variables is greater than the number of observations (p > n). The Least Absolute Shrinkage and Selection Operator (LASSO) estimation method has been proposed in order to obtain more consistent results with the OLS method in the p > n states [1]. The LASSO method regularizes model parameters by shrinking the regression coefficients, reducing some of them to zero so that it is possible to obtain easily interpretable models especially in large data sets. Furthermore, LASSO is a widely preferred method because it provides the shrinkage and variable selection simultaneously. The variable selection part occurs after the shrinkage to be used in the model. The use of LASSO estimator reveals more unbiased models especially when the number of variables is greater than the number of observations.

The LASSO estimator provides strong predictive results in large data sets and is often used in situations with large data sets. In this study, its effect on smaller data sets was examined. The main purpose of this study is to examine the estimation accuracy of the LASSO estimator especially in small data sets, and to compare different information criteria for model selection.

In the simulations, the relationship between the change in the number of observations and the predictive accuracy was investigated basically. For this, GACV and BGACV information criteria were used in addition to the widely used AIC and BIC criteria in model selection. Using different information criteria, the more effective ones among them were determined. Therefore, LASSO estimates were compared according to the criteria by creating the states with different observation numbers for small data sets as the number of dependent variables fixed. And the results were examined.

The rest of the paper is organized as follows. A brief review of the theory of the LASSO and Bernoulli logistic models is described in Section 2. Experiments are presented in Section 3. Results are given in Section 4 and Conclusion part is given in Section 5.

2 Material and Method

In this section the theory of the LASSO and some information criteria used for model selection are mentioned.

2.1 LASSO

Let x_{ij} be the standardized predictors and y_i response values for a linear regression model where $i = 1, \dots, n$ and $j = 1, \dots, p$. The LASSO finds $\beta = \{\beta_j\}$ to minimize in Formula (1).

$$\sum_{\mathrm{i}=1}^{n} \left(y_i - \sum_j x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{1}$$

The parameter $\lambda \geq 0$ controls the amount of shrinkage and regularization [2]. When λ increase, the more coefficients are forced to be zero. Setting $\lambda = 0$ turnes the LASSO estimator into OLS. The LASSO uses the $l1 - penalty$ that yields a convex problem. Convexity is computationally efficient.

2.2 Multivariate Bernoulli Logistic Model

The multivariate Bernoulli distribution as a member of the exponential family is a way to formulate of the binary variables. An important property of that model is the marginal and conditional distributions of a subset of variables follow the multivariate Bernoulli distribution. When the Bernoulli distribution is extended to the multivariate Bernoulli distribution, it is concluded that the results obtained are similar to the logistic regression model. For this reason, when the dependent variable takes binary (0,1) values in the multivariate logistic model, the multivariate Bernoulli logistic regression model expression is used for this model [3].

By adding the LASSO penalty term to the negative likelihood function of the Bernoulli logistic models, the objective function of the multivariate Bernoulli logistic LASSO models can be obtained as in Formula (2).

$$T_\lambda(y,f) = l(y,f) + J_\lambda(f) \tag{2}$$

where $l(y, f)$ is the negative log-likelihood function and $J_\lambda(f)$ is the LASSO penalty term. LASSO estimates can be obtained for bivariate Bernoulli logistic LASSO models by minimizing the target function [4].

3 Experiments

In this study, four different information criteria were used for model selection; Akaike information criterion (AIC) [5], Bayesian information criterion (BIC) [6], Generalized approximate cross validation criterion (GACV) [7] and Bayesian generalized approximate cross validation criterion (BGACV).

The simulations were performed by using the "MVB" package in the R program, where the number of observations (n) for initial beta values was 50, 100, 200, 250, and 300, and the number of dependent variables (k) was 2, 3 and the number of independent variables (p) was 5.

In addition, the repetitive simulations were performed in order to obtain LASSO estimates with AIC, BIC, GACV and BGACV criterion with different number of independent variables for the number of observations 50, 40 and 30 while keeping the number of dependent variables constant for 2 values. Estimation results were obtained by performing 100 repetitions for each case and using AIC, BIC, GACV and BGACV criteria in setting parameter selection. And the initial beta values were determined manually as in Table 1 for generating multivariate Bernoulli simulated data.

Table 1. Initial beta values for k=2

	T=1	T=2	T=3
C1	1,5	0	0
C2	0	-1,5	0
C3	0	0	0
C4	0	0	0
C5	0	0	1

In Table 1, $c^T = (c_0^T, \ldots, C_P^T)$ is the estimated coefficient vector, and beta values are $\beta_0 = (\beta_0^1, \beta_0^2, \beta_0^{12})$,...,$\beta_5 = (\beta_5^1, \beta_5^2, \beta_5^{12})$. The model predictive was obtained as linear predictors in generalized linear models and linear combinations of unknown parameters in linear predictors.

4 Results

The results for the simulations for different numbers of observations and dependent variable were given in Table 2. The results show the percentage value for each criterion of the predicted values obtained in 100 replicate experiments.

Table 2. Accuracy results for p = 5, k = 2, 3, n = 50, 100, 200, 250 and 300

n	k	p	AIC	BIC	GACV	BGACV
50	2	5	84,146	93,902	87,601	88,617
50	3	5	93,870	96,580	92,580	94,870
100	2	5	85,250	95,750	89,166	92,583
100	3	5	94,612	97,645	92,870	94,516
200	2	5	86,833	95,583	90,333	90,750
200	3	5	95,376	97,602	96,989	97,956
250	2	5	84,250	95,583	94,250	96,750
250	3	5	92,419	97,741	97,419	99,677
300	2	5	80,750	95,833	95,166	96,333
300	3	5	93,118	97,634	99,247	99,462

In Table 2, simulation results were given to examine the relationship between the number of observations and predictive accuracy. Therefore, the number of independent variables was kept constant (as p = 5). In addition, to examine the effect of the number of dependent variables on predictive accuracy, the number of dependent variables was determined as two values (as k = 2 and k = 3).

When the results in Table 2 are examined, it was observed that when the number of observations was small (when n = 50 and 100), stronger predictive accuracy were obtained with the BIC criterion. On the other hand, as the number of observations increased, stronger/more accurate estimation accuracy percentages were obtained with the BGACV criterion. Furthermore, the increase in the number of dependent variables also has a positive effect on predictive accuracy.

In order to investigate the effect of the number of independent variables on the predictive accuracy, models were created according to the values of independent variables for the observations 50, 40, and 30 respectively, as the number of observations is fixed.

While the number of dependent variables was fixed to 2, the model was estimated so that the number of observations was 50 and the number of independent variables was 5, 15, 20, 25, 50 and 55, respectively. The results were given in Table 3. As p increased, the BGACV produced more powerful results.

Table 3. Accuracy results for k = 2, n = 50 and p = 5, 15, 20, 25, 50 and 55

n	p	AIC	BIC	GACV	BGACV
50	5	83,750	92,583	85,166	88,00
50	15	60,714	98,809	95,238	90,47
50	20	60,701	95,263	92,807	93,68
50	25	72,777	95,416	97,777	98,33
50	50	94,217	98,639	99,659	100,0
50	55	92,592	93,518	98,765	99,69

The model was estimated so that the number of observations was 40 and the number of independent variables was 5, 10, 20, 40, 45, 50 and 60 respectively, while k=2. The results were given in Table 4. It is shown that BGACV produced more powerful results while p increased.

Table 4. Accuracy results for k = 2, n = 40 and p = 5, 10, 20, 40, 45, 50 and 60

n	p	AIC	BIC	GACV	BGACV
40	5	93,333	96,666	87,500	85,833
40	10	92,592	98,148	95,925	94,444
40	20	83,625	96,198	97,368	97,660
40	40	92,307	97,863	97,435	98,717
40	45	96,212	98,106	96,212	97,727
40	50	94,897	94,897	98,979	99,319
40	60	93,220	98,587	98,870	98,880

The model was estimated so that the number of observations was 30 and the number of independent variables was 5, 10, 15, 30, 35, 40 and 45 respectively, while k=2. The results were given in Table 5. It is shown that BGACV produced more powerful results while p increased for a sample size is fixed.

Table 5. Criteria when k = 2, n = 30 and p = 5, 10, 15, 30, 35, 40 and 45

n	p	AIC	BIC	GACV	BGACV
30	5	90,833	93,333	81,666	77,500
30	10	81,111	95,185	91,111	92,592
30	15	73,809	80,952	94,047	95,238
30	30	86,781	96,551	98,275	97,701
30	35	95,588	93,627	98,039	98,040
30	40	97,008	97,435	98,717	99,572
30	45	96,212	98,106	98,484	98,863

5 Conclusion

As a result; the LASSO estimator is widely used, especially in large data sets, because it yields simpler models and more reliable results. In this study, the LASSO estimator has been studied on smaller data sets and also in p>n states. When dependent variable binary values were obtained, more stable and stronger results were obtained with GACV and BGACV criteria as an alternative to AIC and BIC criteria when LASSO predictive models were obtained. For further studies, the BGACV criterion can be investigated on different samples and models.

Acknowledgment

This work was supported by the Ondokuz Mayıs University (PYO. SCIENCE. 1904.17.002) as the Scientific Research Project.

References

[1] Tibshirani R. “Regression shrinkage and selection via the lasso”. Journal of the Royal Statistical Society. Series B (Methodological), 267-288, 1996.

[2] Tibshirani R. “Regression shrinkage and selection via the lasso: a retrospective”. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73 (3), 273-282, 2011.

[3] Dai B. MVB: Multivariate Bernoulli log-linear model. R package version, 1, 2013.

[4] Dai B. Multivariate Bernoulli distribution models. Technical Report, Department of Statistics, University of Wisconsin, Madison, WI 53706, 2012.

[5] Akaike H. Information theory and an extension of the maximum likelihood principle. Proc. 2nd Inter. Symposium on Information Theory, 267- 281, Budapest, 1973.

[6] Schwarz G. “Estimating the dimension of a model”. The Annals of Statistics, 6 (2), 461-464, 1978.

[7] Xiang D, Wahba G. “A generalized approximate cross validation for smoothing splines with non-Gaussian data”. Statistica Sinica, 675-692, 1996.