A selection procedure using Bootstrap algorithm for multiple regression models

By SUPARMAN

A selection procedure using Bootstrap algorithm for multiple regression models

To cite this article: Suparman 2019 *J. Phys.: Conf. Ser.* **1211** 012029

A selection procedure using Bootstrap algorithm for multiple regression models

Suparman[1]

[1]Department of Mathematics Education, University of Ahmad Dahlan, Indonesia

E-mail: suparman@pmat.uad.ac.id

Abstract. Multiple regression is a mathematical model that is applied in various fields to model the relationship between the dependent variable and several independent variables. If multiple regression is used to model data, then the number of independent variables is unknown. The distribution of error is also unknown. This study proposes a multiple regression model in which the number of independent variables is unknown, and error is distributed arbitrarily. The Bootstrap algorithm is proposed to estimate the number of independent variables, multiple regression coefficients, and error variance. The performance of the Bootstrap algorithm is tested by simulation. The simulation result shows that the Bootstrap algorithm can estimate the number of independent variables, multiple regression coefficients, and error variance well. The bootstrap method can be used to select multiple regression that is suitable for data on the amount of money in circulation. The bootstrap method is also used to determine the factors that affect the amount of money in circulation.

1. Introduction

Multiple regression is a technique to predict the value of a dependent variable from two or more independent variables. Multiple regression is often used to model data in various fields. The multiple regression models are used to predict body fat based on body mass index and body adiposity index [1]. Body fat is a dependent variable. Body period and adiposity index are independent variables. The multiple regression is used to model the formation of haloacetic acid in water [2]. The multiple regression is used to predict weld geometry [3]. The multiple regression models are used to model spatial gait [4]. The multiple regression models are used to examine the relationship between several pairs of geographic pairs, geography, and distance of the environment [5]. The multiple regression is used for modeling cutting performance in the turning process [6]. The multiple regression is used to model rice seed weight as a function of plant height, panicle number, and number of seeds [7].

If multiple regression is used to model data, several problems will arise. First, the number of independent variables is unknown. Second, the multiple regression coefficient is unknown. Third, the error distribution is unknown. In various studies, the number of independent variables is determined and then the regression coefficients are estimated based on data. Regarding error distribution, errors are generally assumed to be normally distributed. In this study, both the number of independent variables and the distribution of errors are assumed to be unknown. The number of independent variables is selected based on the data. Error distributions do not have to be normally distributed but errors are assumed to be arbitrary distributions.

Published under licence by IOP Publishing Ltd

Bootstrap is a method that can be used to estimate the parameters of a mathematical model where errors are distributed arbitrarily. The bootstrap is used to estimate stationary autoregressive moving average (ARMA) model parameters [8]. The bootstrap is used to determine shared trust areas [9]. The bootstrap is combined with subsampling to produce an efficient way of assessing the quality of parameter estimators [10]. The bootstrap is used to model an error in experiments [11]. A residual-based bootstrap is used to detect constant coefficients in the global weighted regression model [12[. The bootstrap is used to build a resampling matrix in a signal [13[. A random weighting method is used for bootstrappingthe critical values on the Monti Portmanteau test [14]. The bootstrap is used in a two-sample test [15]. The bootstrap is used to identify ARMA model orders [16]. The bootstrap is used to determine the Pearson correlation coefficient [17]. The bootstrap is used in the logistic regression model [18]. The bootstrap is used to determine confidence intervals in the field of information systems [19]. The bootstrap is used for errors that are assumed to be not normally distributed [20].

This paper proposes the use of bootstrap to select the number of independent variables and estimate multiple regression coefficients where errors are arbitrary distributions. The structure of this paper is as follows. The first section gives an introduction. The second part explains the method used. The third part describes the results of the research and discussion. The fourth section gives conclusions.

2. Method

Suppose that n states the number of data. For$t = 1, \dots, n$, suppose thaty_t isdependent variable and$x_{t1}, \dots, x_{tk}$arek independent variables ($k = 1, \dots, k_{max}$). Here, k_{max} is the maximum of the number of independent variables. Multiple regression models can be written as:

$$y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + z_t \tag{1}$$

Here,$\beta_0, \beta_1, \dots, \beta_k$ are multiple regression coefficientsandz_t iserror with mean 0 and varianceσ^2.

In matrix form, the equation (1) can be written as:

$$y = x\beta + z \tag{2}$$

where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, x = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \text{and} z = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix}$$

According to the least squares method, the estimator of$\beta$is obtained by minimizing the function

$$\sum_{t=1}^{n} (y - x\beta)'(y - x\beta)$$

This is achieved by partially differentiating$\beta_0, \beta_1, \dots, \beta_k$ and equates the results obtained with 0. This process produces k+1 simultaneous k+1 variable.

$$\begin{gathered} n\beta_0 + \beta_1 \sum_{t=1}^{n} x_{t1} + \cdots + \beta_k \sum_{t=1}^{n} x_{tk} = \sum_{t=1}^{n} y_t \\ \beta_0 \sum_{t=1}^{n} x_{t1} + \beta_1 \sum_{t=1}^{n} x_{t1}^2 + \cdots + \beta_k \sum_{t=1}^{n} x_{t1} x_{tk} = \sum_{t=1}^{n} x_{t1} y_t \\ \cdots \\ \beta_0 \sum_{t=1}^{n} x_{tk} + \beta_1 \sum_{t=1}^{n} x_{tk} x_{t1} + \cdots + \beta_k \sum_{t=1}^{n} x_{tk}^2 = \sum_{t=1}^{n} x_{tk} y_t \end{gathered} \tag{3}$$

In the form of a matrix, equation (3) can be presented as:

$$(x'x)\beta = x'y \tag{4}$$

If the inverse of the matrix $x'x$ exists, by multiplying on both sides of equation (4) with this inverse, it's obtained$(x'x)^{-1}(x'x)\beta = (x'x)^{-1}x'yor\beta = (x'x)^{-1}x'y$. Suppose that$\hat{\beta}$ is an estimator forβthen$\hat{\beta} = (x'x)^{-1}x'y$. Suppose that$\hat{\sigma}^2$ is an estimator forσ^2then$\hat{\sigma}^2 = \frac{y'y-(x\hat{\beta})'y}{n-k-1}$. Statistical criteria C_k is used to estimate the number of independent variables. The value C_k is calculated using the following equation [21]:

$$C_k = \frac{\sum_{t=1}^{n}(y - x\hat{\beta})'(y - x\hat{\beta})}{n} - \frac{2k\hat{\sigma}^2}{n}$$

The best number of independent variables chosen is the number of independent variables that has the smallest value C_k.

Bootstrap[21] is a potential computer-based method for solving statistical inference problems. The basic principle of bootstrap is resampling from the original sample$z_1, z_2, \dots, z_n$. Suppose that$\hat{F}$ is an empirical distribution taken with probability$\frac{1}{n}$ for each observed value$z_1, z_2, \dots, z_n$. The bootstrap samples are defined as random samples of size n arranged from $\hat{F}$. Suppose that B is the number of resampling. The bth bootstrap sample $(b = 1, 2, \dots, B)$ is denoted by$z_1^{(b)}, z_2^{(b)}, \dots, z_n^{(b)}$. The bootstrap samples $z_1^{(b)}, z_2^{(b)}, \dots, z_n^{(b)}$ is a random sample of size n taken with returns from the population $z_1, z_2, \dots, z_n$. The bootstrap data $z_1^{(b)}, z_2^{(b)}, \dots, z_n^{(b)}$ consists of original samples $z_1, z_2, \dots, z_n$ that does not appear, Appears once, twice or more in the resampling process of the original sample.

The computational procedure for determining the bootstrap for parameters k, β dan σ^2 is as follows:

- Calculate $\hat{\beta}$,$\hat{\sigma}^2$and $\hat{C}_k$ based on data y_t and $x_{t1}, \dots, x_{tk}$
- Calculate $\hat{z}_t\,(t = 1, \dots, n)$ using the equation
$$\hat{z}_t = y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{t1} - \cdots - \hat{\beta}_k x_{tk}$$
- For $b = 1, 2, \dots, B$:
    - Do resampling $\hat{z}_t^{(b)}$
    - Calculate $\hat{y}_t^{(b)}$ using the equation
    $$\hat{y}_t^{(b)} = \hat{\beta}_0 + \hat{\beta}_1 x_{t1} + \cdots + \hat{\beta}_k x_{tk} + \hat{z}_t^{(b)}$$
    - Calculate $\hat{\beta}^{(b)}$,$(\hat{\sigma}^2)^{(b)}$ and $\hat{C}_k^{(b)}$
- Calculate $\hat{\beta}_{(boot)}$,$\hat{\sigma}^2_{(boot)}$ and $\hat{C}^{(b)}_{k(boot)}$

3. Results and discussion

As an illustration, we will apply a bootstrap algorithm to determine the point estimation on the synthesis data (simulation case) and real data (case study). The simulation study was conducted to confirm whether the performance of the bootstrap algorithm can work well. The case studies provide examples of the application of theory in solving problems in everyday life.

3.1. Synthesis data

Table 1 shows 20 synthesis data y in multiple regression.

Table 1. Synthesis data.

t	y_t	t	y_t
1	45.2944	11	29.6001
2	51.6638	12	63.69
3	52.7143	13	39.8156
4	46.6236	14	32.7119
5	56.3082	15	41.2902
6	42.858	16	48.6686
7	40.254	17	31.1908
8	27.4063	18	38.7975
9	31.559	19	48.9802
10	38.5711	20	36.8433

The synthesis data above is made using equation (1). The value of k is 2. The values of x_1 and x_2 is presented in Table 2. Errors are made based on a normal distribution with mean 0 and variance 1. The value of k, the regression coefficient and error variance are presented in Table 3.

Table 2. The values of dependent variables.

t	x_{t1}	x_{t2}	t	x_{t1}	x_{t2}
1	3	6	11	3	3
2	4	7	12	4	9
3	7	5	13	6	3
4	3	6	14	2	4
5	7	6	15	3	5
6	7	3	16	4	6
7	6	3	17	3	3
8	1	4	18	8	2
9	4	3	19	6	5
10	9	1	20	7	2

Table 3. The value of the number of independent variables, multiple regression coefficients, and error variance.

k	β	σ^2
2	$(6, 3, 5)$	1

Based on the values of $y, x_1,$ andx_2, the bootstrap method is used to estimate the number of independent variables, multiple regression coefficient, and error variance. Resampling is done as much as $B = 2000$ and $k_{max} = 3$. The estimation of the number of independent variables is done by looking at the statistical value C_k for 3 models (Table 4).

Table 4. The statistical value C_k for 3 multiple regression models.

Dependent variable	Independent variables	k	C_k
y	x_1	1	239.0859
y	x_2	1	109.1284
y	x_1, x_2	2	2.8143

From Table 4 it can be seen that the smallest statistical value C_k is achieved by the 3rd multiple regression equation. Thus, this 3rd multiple regression is the best multiple regression model. So the estimator of k is $\hat{k} = 2$.

Based on the best multiple regression, the parameters of the corresponding multiple regression model are then estimated. The estimation of the number of independent variables, regression coefficients, and error variances is presented in Table 5.

Table 5. The estimated value of the number of independent variables, multiple regression coefficients and error variance.

$\hat{k}_{boot}$	$\hat{\beta}_{boot}$	$\hat{\sigma}^2{}_{boot}$
2	$(5.8995, 3.0213, 5.0453)$	0.9368

If the parameter values and estimator values are both multiple regression coefficients and error variances compared, it can be seen that the difference is relatively small. This shows that the bootstrap algorithm can work "well" in selecting the number of independent variables, multiple regression coefficients, and error variance.

3.2. Real data

Table 6 shows the amount of money in circulation (y) and influencing factors, namely net foreign assets (x_1), net bills to the central government (x_2), bills to central institutions in the form of credit (x_3), and bills to private companies and individuals in the form of credit (x_4) from January 2007 to April 2008 in billion rupiah (source: www.bi.go.id).

Table 6. The amount of money in circulation data and influencing factors.

Month	y	x_1	x_2	x_3	x_4
July 2007	144179	498496	444352	38911	826194
August 2007	149194	498091	443878	40264	846472
September 2007	160327	519360	439649	40281	866979
October 2007	156955	517566	437701	46136	884017
November 2007	161272	518424	447846	44920	908339
December 20007	183419	524703	497478	51038	944074
January 2008	166950	529580	446397	43221	937040
February 2008	165633	543467	433322	40313	955010
March 2008	164995	549049	375976	44748	984424
April 2008	171049	544746	371557	45675	1009072

Based on the values of y, x_1, x_2, x_3andx_4, the bootstrap method is used to estimate the value of the number of independent variables, multiple regression coefficients, and error variance. Resampling is done as much as $B = 2000$ and $k_{max} = 4$. The estimated number of independent variables is done by looking at the statistical value C_k for 15 multiple regression models (Table7).

Table 7. The statistical value C_k for 15 multiple regression models.

Dependent variable	Independent variables	k	C_k
y	x_1	1	1.7714 x 10^8
y	x_2	1	3.2492 x 10^8
y	x_3	1	1.3029 x 10^8
y	x_4	1	1.3118 x 10^8
y	x_1, x_2	2	6.8238 x 10^7
y	x_1, x_3	2	6.8440 x 10^7
y	x_1, x_4	2	1.4807 x 10^8
y	x_2, x_3	2	1.2984 x 10^8
y	x_2, x_4	2	2.8434 x 10^7
y	x_3, x_4	2	7.1298 x 10^7
y	x_1, x_2, x_3	3	9.2289 x 10^7
y	x_1, x_2, x_4	3	4.2453 x 10^7
y	x_1, x_3, x_4	3	7.7687 x 10^7
y	x_2, x_3, x_4	*3*	*2.3552 x 10^7*
y	x_1, x_2, x_3, x_4	4	3.9092 x 10^7

From Table 7 it can be seen that the smallest statistical value C_k is achieved by the 14th multiple regression models. Thus, this 14th multiple regression is the best multiple regression models. This shows that the independent variablex_1 does not affect the dependent variable$y$. Based on the best multiple regression models, the corresponding multiple regression coefficients are then estimated. The estimation of the number of independent variables, multiple regression coefficients, and error variance are presented in Table 8.

Table 8. The estimated value of the number of independent variables, coefficients, and error variance.

$\hat{k}_{boot}$	$\hat{\beta}_{boot}$	$\hat{\sigma}^2{}_{boot}$
3	$(-92898.48, 0.17, 0.52, 0.18)$	7.61 x 10^6

The corresponding multiple regression equation is

$$\hat{y}_t = -92898.48 + 0.17\, x_{2t} + 0.52\, x_{3t} + 0.18 x_{4t}$$

4. Conclusion

The above description indicates that the bootstrap algorithm can be used to produce an estimate parameters in the multiple regression models when the number of independent variables is unknown andthe error has any distribution. The simulation results show that the bootstrap algorithm can estimate the point estimationwell.

The bootstrap method is implemented on the data of the amount of money in circulation (y) and the factors that influence it. The factors that affect the amount of money in circulation are net foreign assets (x_1), net bills to the central government (x_2), bills to central institutions in the form of credit (x_3), and bills to private companies and individuals in the form of credit (x_4) on the month July 2007 to April 2008 (in billion rupiah). By using the bootstrap method, a multiple regression model is obtained, namely:

$$\hat{y}_t = -92898.48 + 0.17\, x_{2t} + 0.52\, x_{3t} + 0.18 x_{4t}$$

This multiple regression model is very useful for decision making, for example to predict the value or calculate the prediction interval of the y variable in the future.

Acknowledgments

The author would like to thank Bank Indonesia for providing data. The author would also like to thank the reviewers who have provided suggestions for improving the quality of this paper.

References

[1] Fuster-Parra P, Bennasar-Veny M, Tauler P, Yanez A, Lopez-Gonzalez A A and Aguilo A 2015 *PLOS ONE* pp 1-14

[2] Ata S, Wattoo F H, Din M I, Wattoo M H S, Qadir M A, Tirmizi S A and Abdullah P 2015 *Arab J. Sci. Eng.* **40** 101-8

[3] Sarkar A, Dey P, Rai R N and Saha S C 2016 *Sadhana* **41** 549-11

[4] Wahid F, Begg R, Lythgo N, Hass C J, Halgamuge S and Ackland D C 2016 *J. of Applied Biomechanics* **32** 128-12

[5] Franckowiak R P, Panasci M, Jarvis K J, Acuna-Rodriguez I S, Landguth E L, Fortin M J and Wagner H H 2017 *PLOS ONE* pp 1-13

[6] Dahbi S, Ezzine L and Moussami H E 2017 *International J. og Engineering Research in Africa* **29** 54-16

[7] Kajonphol T, Seetaput N, Precharattana M and Sangsiri C 2018 *Applied mechanics and materials* **879** 71-7

[8] Kreiss J P and Franke J 1992 *J. of Time Series Analysis* **13** 297-21

[9] Park C 2013 *International J. of Production Research* **51** 4695-9
[10] Kleiner A, Talwalkar A, Sarkar P and Jordan M I 2014 *J. R. Statist. Soc. B* **76** 95-22
[11] Hanson S M, Ekins S and Chodera J D 2015 *J. Comput. Aided. Mol. Des.* **29** 1073-14
[12] Mei C L, Xu M and Wang N 2016 *International J. of Geographical Information Science* **30** 1622-22
[13] Zhen J 2016 *J. of Antennas and Propagation* pp 1-8
[14] Zhu K 2016 *J. R. Statist. Soc. B* **78** 463-23
[15] Olaniran R, Olaniran S F, Yahya W B, Banjoko A W, Garba M K, Amusa L B and Gatta N F 2016 *Annals. Computer Science Series* **14** 46-7
[16] Fenga L 2017 *J. of Probability and Statistics* pp 1-12
[17] Matsumoto T, Murayama Y and Sakatani K 2017 *J. of Human-Computer Interaction* **5** 399-11
[18] Ahmad W M A W, Aleng N A, Ali Z and Ibrahim M S M 2018 *Engineering, Technology & Applied Science Research* **8** 3135-6
[19] Aguirre-Urreta MI 2018*MIS Quarterly* **42** 1001-20
[20] Baokaba T, Korso MNE, Zoubir AM and Berkani D 2018 *Progress in Electromagnetics Research* **81** 125-16
[21] Efron B and Tbshirani R 1993 *An Introduction to the Bootstrap* (New York: Chapman & Hall)

A selection procedure using Bootstrap algorithm for multiple regression models

ORIGINALITY REPORT

3%

SIMILARITY INDEX

PRIMARY SOURCES

1	coqtail.github.io Internet	35 words — 1%
2	"Appendix E", Elsevier BV, 2007 Crossref	14 words — 1%
3	www.slideshare.net Internet	8 words — < 1%
4	"Lake Champlain: Partnerships and Research in the New Millennium", Springer Nature, 2004 Crossref	7 words — < 1%
5	S. N. Lahiri. "Resampling Methods for Dependent Data", Springer Nature, 2003 Crossref	7 words — < 1%
6	"Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)", Springer Nature, 2019 Crossref	7 words — < 1%

EXCLUDE QUOTES	ON	EXCLUDE MATCHES	OFF
EXCLUDE BIBLIOGRAPHY	ON		