Application of Regression Analysis for Small Samples Based on Bootstrap Method

Xin Shibo
School of Economics , Beijing Technology and Business University
BTBU
Beijing, China
e-mail: xsblm@163.com

Ren Bing *
School of Economics , Beijing Technology and Business University
BTBU
Beijing, China
e-mail: searchor@163.com
* Corresponding Author

Abstract—Giving bootstrap method of estimation based on multivariate regression analysis when representation of samples are small, the changing multivariate regression equation by bootstrap method of estimation is used as academic multivariate regression equation, and discussing process of bootstrap samples, repeat numbers, principle of bootstrap method of estimation. By bootstrap samples, giving estimate of regressive coefficients, and evaluating validity of bootstrap method of estimation with binomial regression model, in the end, extending the bootstrap method to non-linear regression model.

Keywords-multivariate regression analysis; bootstrap method; least square method; regressive coefficients; small samples

I. INTRODUCTION

In statistics, regression analysis, based on determined model and data, is generally applied to give the estimated regression coefficients using the least squares method. The effect of regression analysis is excellent to a large extend, but if the selected representative sample was poor, we still directly estimate the regression coefficients using the least squares method, which unavoidably lead to greater estimation error. To solve this problem we introduce bootstrap method, by means of repeated sampling principle, by the deformation model of linear regression analysis, we can configure big data based on a large sample of actual observed data, give the estimated regression coefficients of linear regression model, and promote the bootstrap method to convertible linear model and non-convertible linear model.

The Bootstrap method was first introduced by Professor Efron[1] in 1979, it spread like brush fire in statistical sciences within a couple of decades. This method does not require making assumptions of distribution and adding new sample information and thus to the overall distribution will be a nonparametric method of statistical inference. In statistics, bootstrapping can refer to any test or metric that relies on random sampling with replacement. Bootstrapping allows assigning measures of accuracy to sample estimates. This method allows estimation of the sampling distribution of almost any statistic using random sampling methods. It usually falls in the broader class of resampling methods.

In the event of bootstrap method for nearly 30 years, it has been widely applied to the field of statistics and econometrics, more and more scholars began to focus on bootstrap method, Joel L gives the summary of bootstrap method by way of example, which is about parameter estimation and hypothesis testing based on econometric background, and points out the further study bootstrap method required[2]. Li Fu-chun and Greg T use a bootstrap algorithm to approximate the distribution of the test statistic, and show that the bootstrap distribution converges to the asymptotic distribution of the test statistic in probability[3]. Sílvia G and Maximilien K propose and theoretically justify the application of bootstrap methods for inference in auto regressive panel data models with fixed effects, and give a detailed summary[4]. Bootstrap methods is widely used in various fields, Zuo Guang-hong and Xu Zhao show the stability and self-consistency of CVTrees by performing bootstrap and jackknife re-sampling tests adapted to this alignment-free approach[5]. Akbas and Yusuf E show that there is a weak causal relationship between economic growth and financial development support the neutrality hypothesis in emerging countries, by using the bootstrap panel causality test[6].

In terms of domestic research, Xie Yi-tian and Zhu Yu study on the basic opinion and historical development of bootstrap theory, and they summarize the forefront of research[7]. Zhang Jiang-ling and Zhang Zhong-zhan use Bootstrap method to give a method to calculate the critical value and values, and to explain the process and excellent performance can be achieved through Monte Carlo simulations[8]. Bianling Ou-Bianling and Long Zhi-he apply Bootstrap methods to space correlation Moran's I test of spatial econometric model, MonteCarlo simulation results show that Bootstrap method in most cases is better than the asymptotic test from the perspective of efficacy[9]. Ding Xian-wen and Zou Shu research on the Bootstrap method for calculating confidence intervals of normal distribution and Poisson distribution, through simulation, they make recommendations about calculating confidence intervals campared to the bootstrap method and classical method[10]. Liu Wei and Chang Zhen-hai conduct overall statistical functional estimation by using Bootstrap method, the results show that the Bootstrap method works better in discrete distribution. In a continuous distribution based on bootstrap, when the sample size n is less than or equal to 5, the symmetric distribution is better than asymmetric distribution, while the asymmetric distribution is better when the sample size n is more than or equal to 6[11]. In the

applications, Xie Shui-yuan and Liu Yuan, through the inspection of FDI and Chinese Economic Growth, verify the bootstrap method provides an effective analysis method for the economic measurement model when the distribution of the error term is unknown[12]. Duan De-feng, Wang Jian-hua and Song Hong-fang propose a risk measurement model based on Bootstrap method, the results show that if the sample was small, this method can better manage credit risk[13].

II. THE PRINCIPLE OF BOOTSTRAP METHOD BASED ON LINEAR REGRESSION

A. The Deformation of Linear Regression Model

Multiple Linear Regression Model can be described as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \qquad (1)$$

In which y is the dependent variable, $x_1, x_2, \cdots, x_k$ are the independent variables,k is the number of independent variables. Assuming that n samples was given as $(y_i, x_{1i}, \cdots, x_{ki})$, $i = 1, \cdots, n$, we substitute (1) with $(y_i, x_{1i}, \cdots, x_{ki})$, $i = 1, \cdots, n$ as follows:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} , \ i = 1, \cdots, n \qquad (2)$$

The n equations in (2) are added up and divided by n,the answer is

$$\overline{y} = \beta_0 + \beta_1 \overline{x_1} + \beta_2 \overline{x_2} + \cdots + \beta_k \overline{x_k} \qquad (3)$$

In (3), $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\overline{x_m} = \frac{1}{n}\sum_{i=1}^{n} x_{mi}$, $m = 1, \cdots, k$, the regression coefficients of (1) and (3) are identical, only independent variable and the dependent variable changes, (3) is the theoretical model of multiple regression analysis using bootstrap method.

B. The Sample of Bootstrap Method

Efron's research conclude such a conclusion: if the sample size of the original sample is n, then non-duplicate sample group number is C_{2n-1}^{n} by repeating the sampling sample remains available n, for simplicity of description denote $B = C_{2n-1}^{n}$, Table 1 shows the different samples n corresponding to non-repeated sample groups based on bootstrap method .

TABLE I. THE DIFFERENT SAMPLES N CORRESPONDING TO THE GROUPS OF B BASE ON BOOTSTRAP

n	2	3	4	5
$B = C_{2n-1}^{n}$	3	10	35	118
n	6	7	8	9
$B = C_{2n-1}^{n}$	462	1716	51480	1750320

As can be seen from Table 1, when the law does not duplicate sample buffet obtained by the bootstrap method theoretically has reached a large sample volume statistics mentioned in the question, that sample using the least squares method of parameter estimation Bootstrap sample requirements are fully meet.Now given sample generating step bootstrap method, assuming the original sample size n, with the help of Matlab7.8[14]:

Step1. Generate n uniformly distributed random numbers $z_1, z_2, \cdots, z_n$ and calculate $[z_i \times n] + 1$, $i = 1, \cdots, n$ as the basis of selection in step 2 of the original sample.

Step2. According to the random numbers in the step1, to determine a set of bootstrap samples $(y_{[z_i \times n]+1}, x_{1\,[z_i \times n]+1}, x_{2\,[z_i \times n]+1}, \cdots, x_{k\,[z_i \times n]+1})$, $i = 1, \cdots, n$, and calculate their averages recorded as $(\overline{y_1^*}, \overline{x_{11}^*}, \overline{x_{21}^*}, \cdots, \overline{x_{k1}^*})$.

Step3. Repeat step2 B times, then get B sets of bootstrap samples $(\overline{y_1^*}, \overline{x_{11}^*}, \overline{x_{21}^*}, \cdots, \overline{x_{k1}^*})$, $(\overline{y_2^*}, \overline{x_{12}^*}, \overline{x_{22}^*}, \cdots, \overline{x_{k2}^*})$, ... , $(\overline{y_B^*}, \overline{x_{1B}^*}, \overline{x_{2B}^*}, \cdots, \overline{x_{kB}^*})$.

According to the samples $(\overline{y_b^*}, \overline{x_{1b}^*}, \overline{x_{2b}^*}, \cdots, \overline{x_{kb}^*})$, $b = 1, 2, \cdots, B$. (3) is changed into

$$\overline{y_b^*} = \beta_0 + \beta_1 \overline{x_{1b}^*} + \beta_2 \overline{x_{2b}^*} + \cdots + \beta_k \overline{x_{kb}^*} \qquad (4)$$

Use the least squares method to estimate the parameter of the regression coefficient on the basis of (4), and then get the estimator $\beta_0, \beta_1, \cdots, \beta_k$ of the regression coefficient $\beta_0, \beta_1, \cdots, \beta_k$. The specific form is similar to the results acquired from (2). Here is not repeated.

C. The Essence Of Linear Regression Based on Bootstrap Method

The bootstrapping samples presented in this paper is using the repeated sampling ideas of bootstrap. To get the same samples as original sample size every time, and then take the averages as a sample of parameter estimation using the least squares method based on (3). From the constructing process of samples, you can see that only n groups of values are true in the new samples acquired from the bootstrap. (i.e. each extraction are the same samples, repeated n times), other samples are pseudo samples which is conform to the type of (3). at the same time ,for the processing method is only average, the value of bootstrapping samples must between the minimum and maximum of true values. To draw a conclusion, the basic principle of linear regression method of buffet is based on the form of the straight line of actual samples (3),obtain the bootstrapping samples which is between the most values and conform to the equation of the straight line, and further to increases the sample size of parameter estimation and improve the accuracy of parameter estimation.

D. The Discussion Of Repeat Time

It Had been given in section 2.1, if the original sample size is n, the number of bootstrapping sample groups whose size is n by repeated sampling is, it need to be repeated B times in theory, but according to the steps of generating bootstrapping samples in section2.2, the repeated steps of B times may bring about repetition. So it need to be further discussed whether it should repeat B times or not . This section takes single-element regression model that is based on small sample as an example, the model is based on (4), and provide the comparison of regression results when the sample sizes are 4, 5 and 6, while the corresponding repetition times are $B = C_{2n-1}^{n}$, 2B, 3B separately. The data is from the example 1-4 of last 4 to 6 years in literature [15] . the reason to choose is that this study is intended to explore the repetition times in order to get stable parameter estimate using bootstrap.

Therefore , the size of the actual sample size has no effect on the determination of repetitions times. The results based on the SPSS is shown in table 2.

TABLE II. COMPARISON OF THE REGRESSION EFFECTS UNDER DIFFERENT SAMPLE SIZES AND REPETITIONS

Sample Size / repetitions		B	2 B	3 B
4	Coefficient	0.981426	0.982907	0.987783
	Sig.	0.0000	0.0000	0.0000
	R-square	0.985772	0.98 6313	0.988684
	Adjusted R-square	0.985341	0.986112	0.988574
5	Coefficient	1.004972	1.011051	1.012834
	Sig.	0.0000	0.0000	0.0000
	R-square	0.994004	0.992074	0.993888
	Adjusted R-square	0.993952	0.992015	0.993871
6	Coefficient	1.005101	1.032193	1.040145
	Sig.	0.0000	0.0000	0.0000
	R-square	0.995321	0.994912	0.995214
	Adjusted R-square	0.994872	0.994102	0.994917

According to the results of table 2, it can be seen from the coefficient of regression, acceptance probability, goodness of fit and adjustment coefficient of goodness-of-fit, under the condition of the same sample size, when the repetition times is from B to 2 B, 3 B, the four indexes have almost no differences, so the repetition times can be .

III. THE APPLICATION OF LINEAR MODEL AND FITTING EFFECT TEST

A. The Fitting Steps of Regression Analysis

In practice, we always tend to take correlation analysis to determine whether there is any internal correlation relationship between the variables before the regression analysis. The representative quality of the samples, however may have overwhelming influence on the description of the true intrinsic relations between variables. the solutions to the problem are followed.

If the determined variables fulfill formula (1), then it will also satisfy formula (3). the bootstrapping samples by the means of bootstrap, will also fulfill formula (4). The correlation analysis we used to study the intrinsic relations between variables are based on the bootstrap samples. Then we build a correlation matrix on the basis of B in size bootstrap samples, to study the correlation analysis. If the results show little or none correlation relations, it indicates that the variables are inappropriate. While, the results show a strong correlation relation, we can set up formula (4) and accordingly start the process of parameter estimation and test. If the variables meet the standards, it means strong correlation relations between research variables. Linear regression equation are obtained , if not ,it means non-existent correlation relations between variables, which requires other nonlinear regression method to fit. In summary, the steps of regression analysis are followed:

1.Determine the bootstrap method sample $(\overline{y_b^*}, \overline{x_{1b}^*}, \overline{x_{2b}^*}, \cdots, \overline{x_{kb}^*})$, $b = 1, 2, \cdots, B$

2. Establish correlation matrix, and proceed correlation analysis.

3. Establish multiple linear regression equation based on formula (4) estimate the parameters, and then judge the test outcome.

4. If the test results are good, take the to-be-estimated parameter into formula (1) and proceed the fitting effect test based on the actual data, in terms of formula (1) as theory model. If the fitting results are good, we can use it to analyze and predict, otherwise, it needs other models to fit.

B. The Fitting Effect Test of Bootstrap Method

The data in this section is from the example 2-1 of literature [15], the result of regression analysis based on (1) in literature [15] is

$$y = 20.069 + 35.456x_1 + 10.872x_2$$

$$R^2 = 0.946846 \quad \overline{R^2} = 0.939253$$

Std.Error(β) (10.04447) (0.972182) (5)

t Prob. 0.0030 0.0000

F Prob. 0.0000

The result based on the bootstrap estimation of regression model is

$$\overline{y} = 19.47017 + 35.47144\overline{x_1} + 10.88005\overline{x_2}$$

$$R^2 = 0.956965 \quad \overline{R^2} = 0.956929$$

Std.Error(β) (0.686682) (0.066302) (6)

t Prob. 0.0000 0.0000

F Prob. 0.0000

It can be seen from the index Std.Error(β) ,the fluctuation of (6) is smaller, the method we give in this paper is better, but (6) is based on (4), R^2 in (5) and $\overline{R^2}$ in (6) has no comparability, (6) can not show the pros and cons of fitting effect of actual problem, take the regression coefficient of (6) into the model of (1),and calculate the R^2 and $\overline{R^2}$ based on the actual data, the result is $R^{*2} = 0.947584$ and $\overline{R^{*2}} = 0.940096$,this fitting effect is almost no difference with the result of traditional regression method. But the estimate parameters of estimate from (6) is more stable than (5) , which means the result based on (6) is better.

C. The Evaluation of Fitting Effect.

The Least-square method is based on the actual observation data, which is a kind of method to obtain the parameter estimates by minimize the sum of the practical and theoretical error ,so the goodness of fit based on the actual data should be the best, there will be no other ways to transcend, and using the bootstrap estimation method of multiple regression presented in this paper estimation is better than using least-squares method is because bootstrap is based on the bootstrapping samples which is obtained from the initial samples. It increased the actual sample size of parameter estimator, further to improve the estimation precision of parameter, and then use the estimated parameters to calculate the actual goodness of fit, the result must be better than using the least-square method.

IV. THE EXTENSION OF NONLINEAR MODEL APPLICATION

A. The Convertible Linear Model.

This section here takes the binary exponential curve model as an example, given calculation methods and procedures of Bootstrap estimation of nonlinear models.

Taking the natural logarithm of formula on both sides of the binary exponential curve model $y = \beta_0 \beta_1^{x} \beta_2^{z}$:

$$\ln(y) = \ln(\beta_0) + \beta_1 \ln(x) + \beta_2 \ln(z) \quad (7)$$

If the sample is (y_i, x_i, z_i) ($i = 1, 2, \cdots, n$), you can obtain the natural logarithm of the sample $(\ln(y_i), \ln(x_i), \ln(z_i))$, $i = 1, 2, \cdots, n$, and then put the natural logarithm of the sample $(\ln(y_i), \ln(x_i), \ln(z_i))$ ($i = 1, 2, \cdots, n$) into (7), averaging formula (8):

$$\overline{\ln(y)} = \ln(\beta_0) + \beta_1 \overline{\ln(x)} + \beta_2 \overline{\ln(z)} \quad (8)$$

In the formula (8), $\overline{\ln(y)} = \frac{1}{n}\sum_{i=1}^{n} \ln(y_i)$, $\overline{\ln(x)} = \frac{1}{n}\sum_{i=1}^{n} \ln(x_i)$, $\overline{\ln(z)} = \frac{1}{n}\sum_{i=1}^{n} \ln(z_i)$, $i = 1, 2, \cdots, n$.

Take formula (8) as a new regression equation, and do repeated sampling among the natural logarithm of the sample, then we can get the B sample size n sample buffet law, according to the least squares method to obtain the estimated regression coefficients.

B. The Non-convertible Linear Model

The Taylor expansion, a conventional method to process the non-convertible linear model, means solely taking the first-order linear part of The Taylor expansion, and by that, using the least square method to estimate the regression coefficients. The nonconvertible linear model, whose ultimate method to estimate the regression model is the multivariate linear regression analysis as well, is actually approximate to linear approximation, therefore, we can easily come to the conclusion that it is theoretically possible that the bootstrap method applies to the non-convertible linear model without any further discussion.

V. CONCLUSIONS AND SUGGESTIONS

Bootstrap method will be applied to multiple regression analysis, it can solve the problems when the samples lack representations, compared with traditional regression parameter estimation method bootstrap method is more effective. But through the process of sample construction based on bootstrap method , we can see it also has a certain dependence to the sample, maximum and minimum values at the maximum and minimum values chosen can't close to the actual problem, Using Bootstrap sample obtained still do not have better representation,, continued regression analysis using the bootstrap method is also not a good fit results, so if known minimum and maximum practical problems can be extracted small sample size regression analysis using bootstrap method to save sampling costs; on the contrary, in the economic and other under objective conditions permit, it is recommended to use the sample still large sample size conduct regression analysis, but in the specific process that regression coefficient is calculated, the principles and ideas of bootstrap can be still used,in order to improve the accuracy of parameter estimation and forecasting results.

ACKNOWLEDGEMENT

Finally we extend our sincere appreciation to Beijing Talents Funded Projects(Serial number: 2011D005003000011), Talents Teach-Youth Fellowship Program(Serial number: YETP1458), Beijing Social Science Foundation Research Base Project(Serial number: 15JDJGB076) and Circulation capital Beijing Philosophy and Social Sciences Research Base.

REFERENCES

[1] Bradley E, Robert J. An Introdution to the Bootstrap [M]. Chapman&Hall, 1993.

[2] Joel L. The Bootstrap in Econometrics [J]. Statistical Science. 2003, 18(2):211-218.

[3] Li Fu-chun, Greg T. A consistent bootstrap test for conditional density functions with time-series data [J]. Journal of Econometrics. 2006, 133(2):863–886.

[4] Sílvia G, Maximilien K. Bootstrap inference for linear dynamic panel data models with individual fixed effects [J]. Journal of Econometrics. 2015, 186(2):407-426 .

[5] Zuo Guang-hong, Xu Zhao, Yu Hong-jie. Jackknife and Bootstrap Tests of the Composition Vector Trees [J]. Genomics Proteomics & Bioinformatics. 2010, 8(4):262-267.

[6] Akbas, Yusuf E. Financial development and economic growth in emerging market: bootstrap panel causality analysis [J]. Theoretical & Applied Economics. 2015, 22(3):171-186

[7] Xie Yi-tian, Zhu Yu. Historical development and advanced research of Bootstrap Methods [J]. Statistics & Information Forum,2008, 02:90-96.

[8] Zhang Jiang-ling, Zhang Zhong-zhan, Zhang Sai-yin. Multi-Sample Testing Based on Bootstrap Method for Simple Stochastic Ordering [J]. Journal of Systems Science and Mathematical Sciences,2014, 08:950-959.

[9] Ou Bian-ling, Long Zhi-he, Lin Guang-ping. Simulation Analysis of Bootstrap Test Efficacy in Spatial Econometric Model.Journal of Nanjing University of Science And Technology [J] ,2010, 11:155-160.

[10] Ding Xian-wen, Zou Shu,Lin Jin-guan. Comparison between Bootstrap Method and Classical Method in Interval Estimation [J]. Statistics and Decision,2012, 23:72-73.

[11] Liu Wei, Chang Zhen-hai. Comparative Research of Statistical Functional Estimation Based on Bootstrap Method [J]. Statistics and Decision,2013, 02:71-72.

[12] Xie Shui-yuan, Liu Yuan. The Application of Bootstrap Method in Economic Analysis [J]. China Market,2011, 23:211-212.

[13] Duan De-feng, Wang Jian-hua, Song Hong-fang. Measure of Credit Risk Based on Bootstrap [J]. Journa of WuhanUniversity of Technology,2011,02:328-330.

[14] Zhang De-feng. Probability and Mathematical Statistics Based on Matlab [M].Beijing:China Machine Press. 2010.

[15] Li Bao-ren, Qiao Yun-xia,Wang Qin-ying. Econometrics [M]. Beijing:China Machine Press. 2008.