International Journal of Current Advanced Research
ISSN: O: 2319-6475, ISSN: P: 2319-6505, Impact Factor: 6.614
Available Online at www.journalijcar.org
Volume 9; Issue 11 (A); November 2020; Page No.23271-23273
DOI: http://dx.doi.org/10.24327/ijcar.2020.23273.4609

Research Article

BOOTSTRAPPING METHODS AND SOME ASPECTS

K. Balaraju, Jakkula Srinivas and A. Rajendra Prasad

Department of Statistics, Kakatiya University, Warangal, (T.S), India.

ARTICLE INFO

Article History:
Received 24th August, 2020
Received in revised form 19th September,, 2020
Accepted 25th October, 2020
Published online 28th November, 2020

Key words:

Bootstrapping, Estimation, Monte Carlo method, Sampling distribution, Testing of Hypothesis, Test statistic.

ABSTRACT

This article offers some explanations for the phenomenon and examples for understanding of "Bootstrapping". The purpose of this article is to defend the view expressed in introduction and prescribe some suggestions from an unexpected but useful source "the bootstrap" where other methods fail to yield required results. We begin by explaining the concept of sampling distribution. After exposing the bootstrap, some examples illustrate how bootstrap exercises can promote understanding of the sampling distribution concept and efficiently useful in predicting the results.

INTRODUCTION

The concept of sampling distribution set the ground rules for the game of Statistics. Most of what statisticians do is either finding "good" estimates for unknown parameters or testing of hypotheses. The sampling distribution creates the following underlying logic for these two activities.

Using a formula β^* to produce an estimate of β can be conceptualized as the statistician shutting his or her eyes and obtaining an estimate of β by reaching blindly into the sampling distribution of β^* to obtain a single number. Here β could be a parameter or a value to forecast.

Because of this, choosing between β^* and a competing formula β^{**} comes down to the ambiguity that would we prefer to produce the estimate of β by reaching blindly into the sampling distribution of β^* or by reaching blindly into the sampling distribution of β^{**}.

Because of above, desirable properties of an estimator β^* are defined in terms of its sampling distribution. For example, β^* is unbiased if the mean of its sampling distribution equals the number β being estimated. This explains why statisticians spend so much algebraic energy figuring out sampling distribution properties, such as mean and variance.

The properties of the sampling distribution of an estimator β^* depend on the process generating the data. So an estimator can be a good one in one context but a bad one in another. When we move from one textbook topic to another we are moving from one data generating process to another, necessitating a re-examination of the sampling distribution properties of familiar estimators and development of new estimators designed to have "better" sampling distribution properties.

Test statistics have sampling distributions. When we test hypotheses we carefully choose a test statistic which, if the null hypothesis is true, has a sampling distribution described by numbers we know. Many such test statistic sampling distributions are described by tables in the back of statistics text books. The explanation of how and why we accept or reject a hypothesis is built on the logic of the sampling distribution. If we understand sampling distribution, the rules of the game, we will understand the logic of hypothesis testing.

What Is Bootstrapping

Only in simple cases theory in statistics deduce the properties of a statistic's sampling distribution. In most cases theory is forced to use asymptotic algebra, producing results that apply only when the sample size is very large. Although in many cases these asymptotic results provide remarkably good approximations to sampling distributions associated with typical sample sizes, one can never be sure. Because of this, statisticians have turned to the computer to discover the sampling distribution properties of statistics in small samples, using Monte Carlo method.

In the Monte Carlo method the computer is used for the data generation, creating several thousand typical samples, calculating for each sample the value of the statistic required and then using these thousands of values, characterize the

Corresponding author:* **K. Balaraju

statistic's sampling distribution by estimating its mean, variance and mean square error or any other required property.
In data generating process the computer needs to forecast errors from an error distribution, which in reality is unknown. For convenience, in most Monte Carlo studies errors are drawn from a normal distribution. But in many problems the reason we believe that the asymptotic results are not reliable in small samples, as we do not believe that the errors are distributed normally. In such cases traditional Monte Carlo methods do not produce effective results. To deal with this problem, we must find a way of drawing errors more representative of the unknown actual error distribution. **Bootstrapping is a method for solution of this problem.**

Bootstrapping Versus Monte Carlo

Bootstrapping is a variant of Monte Carlo in which the error distribution from which the computer draws errors is an artificial distribution with equal probability on all of the residuals from the initial estimation of the model under investigation. This is typically described as randomly drawing with replacement from the set of ordinary least squares residuals. In effect the actual, unknown distribution of errors is being approximated by this artificial distribution. This bootstrapping procedure has been shown to perform remarkably well. It produces estimates of sampling distributions of statistics that are surprisingly accurate and so has become increasingly popular in statistical analysis.
In particular, "Bootstrapping" is used for two main purposes.

Bootstrapping used in testing

Suppose one is using an 'F' test to test some hypothesis, but because he fears that his problem is characterized by non-normal errors he is worried that for his modest sample size the sampling distribution of F- statistic under the null hypothesis is not accurately characterized by the figures in the F- table . In particular one might fear that the 5 percent critical value found in this table may for his problem be more like a 25 percent critical value!

By bootstrapping this F-statistic under the null hypothesis, he can create a description of the F-statistic's sampling distribution suitable to his problem. By using this distribution instead of the tabled F-distribution he can choose the critical value to ensure that the resulting type-I error is indeed 5 percent.

Bootstrapping for estimating Confidence Intervals:

Suppose one is forecasting a variable and wish to produce a confidence interval for the forecast. The usual way of calculating such a forecast interval is, to find the standard error of the forecast, multiply it by a suitable critical value taken from the t - distribution and then add and subtract the result from his forecast. This procedure could be very inaccurate, however, for several reasons. One can also use a search procedure to develop some specifications, the variable being forecast may be a nonlinear function of parameters estimated via these specifications, and the errors may not be distributed normally. It is not known how to find the standard error of such a forecast, and even if it was, the forecast would surely not have a t - distribution, nor be symmetric. By bootstrapping this entire estimation procedure the actual sampling distribution of the forecast could be estimated, allowing an appropriate confidence interval to be produced.

We here illustrate some examples to elucidate use of bootstrapping in various situations.

Example

Suppose we have 25 observations (say) on variables Y and X and assume that Y = a + bX + e, where the classical linear regression (CLR) model holds (but not the classical normal linear regression model, which implies that the errors are distributed normally). One can proceed with ordinary least squares and obtain estimates 0.50 and 2.25 of 'a' and 'b', with corresponding estimated variances 0.04 and 0.01 (say), saving the residuals in a residual vector 'res'. One can develop a computer program in any language for the following algorithm.

i. Draw 25 'e' values randomly with replacement from the elements of 'res'.
ii. Compute 25 values of Y using: y=0.5 + 2.25*x + 1.043*e.
iii. Regress 'y' on 'x', obtaining an estimate $\hat{\beta}$ of β and its standard error 'se'.
iv. Compute t = $\{(\hat{\beta} - \beta)^2\}$/se and save it.
v. Repeat from (i) to obtain 2000 values of't'.
vi. Arrange these t - values in ascending order.
vii. Print the 50th t - value t[50] and the 1950th t - value t[1950].

Then:

a. Explain what this program is designed to do.
b. The answer should be: This program is producing 2000 values of't', those can be used to estimate the sampling distribution, under the null hypothesis, β=2 of the t - statistic for testing β = 2.
c. Suppose t[50] = -2.634 and t[1950] = 2.717 ,what conclusion would you draw?
d. The answer should be: The t–value obtained from the actual data is 2.50 because it lies within the two-tailed 5 percent critical values of -2.634 and 2.717, we fail to reject the null hypothesis at this significance level.

Example

Suppose we have 28 observations on 'y', 'x', and 'z' and let y = α + βx + γz + ε, where the CLR model holds. We run ordinary least squares and obtain estimates 1.0, 1.5 and 3.0 of α, β and γ saving the residuals in a vector 'res'. We have to program a computer to do the following:

i) Draw 28 values of ε randomly with replacement from the elements of 'res'.
ii) Compute 28 values of y using: y=1.0 + 1.5*x + 3.0*z + 1.058* ε .
iii) Regress 'y' on 'x' and 'z' obtaining an estimate $\hat{\beta}$ of 'β' and $\hat{\gamma}$ of 'γ'.
iv) Compute: r = $\hat{\gamma} / \hat{\beta}$ and save it.
v) Repeat from (i) to obtain 4000, values of 'r'.
vi) Compute and print the average (av) of the 'r' values and their variance (var).
vii) Arrange these 'r' values in ascending order.

Then:

i. An estimate of the bias of $\hat{\gamma}/\hat{\beta}$ as an estimate of γ/β is estimated by av $^{-2}$.

ii. An estimate of the standard error of $\hat{\gamma}/\hat{\beta}$ is estimated by square root of 'var' .

iii. Testing the null hypothesis that the bias = 0 is a t - statistic for testing the null hypothesis is the estimated bias divided by the square root of its estimated variance.

d) A 90 percent confidence interval for γ/β is given by the interval between the 200th and the 3800th 'r' values, adjusted for any bias.

Example

Suppose that the CLR model applies to $y = a + \beta x + e$, except that the error variance is larger for the last half of the data than for the first half. Let the error be not distributed normally. In this case, Goldfeld-Quandt statistic will not have an F - distribution for the given sample size. Given some data, we can able to explain how to bootstrap the Goldfeld-Quandt statistic to test the null hypothesis that the error variances are the same.

Example

Suppose we have programmed a computer to do the following:

i) Draw 25 'x' values from a uniform distribution between 4 and 44.
ii) Set ctr = 0.
iii) Draw 25 values from a standard normal distribution and multiply all the negative values by 9 to create 25 'e' values.
iv) Compute 25 values of y as y=3 + 2*x + e.
v) Regress 'y' on 'x', saving the intercept estimate as 'int', the slope estimate as 'b', the standard error of 'b' as 'se' and the residuals as a vector 'res'.
vi) Compute $t=b^2/se$ and save it.
vii) Compute 25 values of y as y= int + 2*x + 1.043*b*e, where b*e is drawn randomly with replacement from the elements of 'res'.
vii) Regress 'y' on 'x' and compute $bt(1) = b^2/se$, where 'b' is the slope coefficient estimate and 'se' is its standard error.
viii) Repeat from (vii) to obtain 200 values of 'bt'.
ix) Arrange these 'bt' values in ascending order.
x) Add one to 'ctr', if t is greater than the 190th ordered 'bt' value.
xi) Repeat from (iii) to obtain 500 values of t.
xii) Calculate the fraction of these t values.

The above program is designed to compare the actual type-I error of a traditional one-sided t -test and its bootstrapped version. The context is a linear regression in which the error terms have come from an asymmetric distribution, a nominal significance level of 5 percent has been employed, the null hypothesis is that the slope= 2, and the alternative hypothesis is that the slope > 2. The tabled 5 percent critical value for the t - distribution with 23 degrees of freedom is 1.56.

DISCUSSION AND CONCLUSIONS

A sampling distribution reflects relative frequencies with which different values of a statistic would be obtained if different errors had been drawn.

It is possible to have techniques for bootstrapping when the errors are not exchangeable. An example is heteroscedasticity problem associated with an explanatory variable, causing large (in absolute value) errors to be more likely to be attached to some observations than others. A very different bootstrap resampling procedure is used to deal with this, in which a bootstrapped sample is formed by drawing with replacement from the set of original observations (where each dependent variable value and its associated independent variables values is a single observation). For some applications the bootstrapped samples must be created in imaginative ways as well.

The number of bootstrapped samples required varies from case to case. Efron [1] suggests that estimation of bias and variance requires only about 200 but estimation of confidence intervals, and thus use for hypothesis testing, requires about 2000.
For bootstrapping a pivotal statistic would require bootstrapping a t - statistic (where the standard error is calculated using an asymptotic formula) for $\hat{\gamma}/\hat{\beta}$ (Example 4.2). The confidence interval would be calculated by taking the bootstrap critical t - values and multiplying them by the bootstrapped estimated standard error. The main message of this article is that to learn the sampling distribution concept and that the way to do this is to provide sample examination questions that require them to demonstrate. An advantage of this is that bootstrapping is becoming common in applied statistics.

References

1) Efron, B., "Better bootstrap confidence intervals. Journal of the American Statistical Association", 85(397): 79-89 (1987).
2) Garfield, J., "How students learn statistics. International Statistical Review", 63(1):25-34 (1995).
3) Simon, J. L., and P. C. Bruce., "Resampling: A tool for everyday statistical work.",Chance 4(1): 22-32 (1991).
4) Zerbolio, D. J., "A 'bag of tricks' for teaching about sampling distributions.", Teaching of Psychology 16(2): 207-209(1989).

How to cite this article:

K. Balaraju (2020) 'Bootstrapping Methods and Some Aspects ', *International Journal of Current Advanced Research,* 09(11), pp. 23271-23273. DOI: http://dx.doi.org/10.24327/ijcar.2020.23273.4609

****