Statistic Corner

A synopsis of resampling techniques

Alessandro Brunelli

Department of Thoracic Surgery, St. James's University Hospital, Leeds, LS9 7TF, UK

Correspondence to: Dr. Alessandro Brunelli, MD, Consultant Thoracic Surgeon LTHT, Honorary Senior Lecturer University of Leeds. Department of Thoracic Surgery, St. James's University Hospital, Bexley Wing, Beckett Street, Leeds, LS9 7TF, UK. Email: brunellialex@gmail.com.

Abstract: Bootstrap is a computer intensive technique of resampling with replacement, which can be applied in many statistical analytical tests. The article describes the most frequent situations where bootstrap resampling can be applied in thoracic surgical research: variable selection for multivariable regression analysis, internal validation of regression equations, model validation. Practical examples for programming bootstrap in commercially available statistical software are finally reported.

Keywords: Resampling statistics; bootstrap; risk modelling; thoracic surgery

Submitted Aug 25, 2014. Accepted for publication Sep 04, 2014.
doi: 10.3978/j.issn.2072-1439.2014.09.09
View this article at: http://dx.doi.org/10.3978/j.issn.2072-1439.2014.09.09

Introduction

Re-sampling statistics has been only recently popularized in biomedical science owing to the availability of statistical programs and computing capability.

Bootstrap is perhaps the most popular of these computer intensive methods in our specialty.

Bootstrap refers to a simulation technique proposed by Efron and colleagues more than 30 years ago (1,2). It consists of generating observations from the distribution of the original sample of patients at hand.

Each simulation results in a new sample typically of the same size (number of individuals) as the original. The simulated sample is generated through a process of random selection (with replacement) of individuals from the original sample.

Sampling with replacement means that at each step of the simulation every individual from the original sample (or dataset) is again eligible to be selected, irrespective of whether he has already been selected.

Therefore, in each bootstrap sample some of the original individuals may not be represented and others may be represented more than once. Hypothetically, a bootstrap sample can be composed by a population represented by the same individual randomly sampled n times.

This random sampling with replacement is repeated to generate hundreds or thousands (typically 1,000) new simulated populations (samples) ensuring accurate statistics without assumptions by combining and analyzing the information generated from these many datasets.

The name "bootstrap" derives from the expression "pulling yourself up by your own bootstraps," reflecting the fact that one could develop all the statistical testing necessary directly from the actual data at hand.

These techniques have been applied to all analytical processes. However, there are situations when bootstrap resampling has been utilized more often in our field of scientific research.

When do I apply bootstrap in my research?

Selection of variables for multivariable analysis

Linear or non-linear regression analyses require the selection of variables to be entered in the model. One of the most used methods to screen variables of interest for multivariable regression analysis is univariable testing.

Usually those variables with a pre-determined P value (typically $P<0.05$ or $P<0.1$) are selected and used as independent variables in the multivariable analysis. However, one has to pay attention to the correlation of the numeric variables to be entered in the regression model. Highly correlated variables should not be entered simultaneously in the same regression iterations as they

will cause a problem of multicollinearity that will affect the results of the analysis. To obviate this problem one can perform multiple regression iteration using one or the other correlated variable at each time and then select the best performing model. Another method is to select the independent variable among those with a high correlation (r>0.5) using bootstrap resampling simulation. This is the method I prefer when facing this problem.

Basically, the univariable comparative analysis (unpaired *t* Student or Mann Whitney test) is repeated in the 1,000 simulated samples generated with bootstrap bagging. The variable that will result associated with outcome (P<0.05) in more samples will be the one to be selected and entered in to the multivariable analysis.

A practical example is the selection of variables such as forced expiratory volume in 1 second (FEV1), forced vital capacity (FVC), FEV1/FVC ratio or predicted postoperative forced expiratory volumes in 1 second (ppoFEV1) for a logistic regression analysis to build a predictive model of in-hospital mortality after lung resection. All these variables are highly correlated each other (r>0.5) and often associated with mortality at univariable analysis. They cannot be used together in the same regression iteration. Bootstrap is my method of choice to verify which one of those variables is most frequently associated with outcome when tested in 1,000 simulated samples.

To this purpose I use the following programming syntax in the Stata 12 statistical software (Stata Corp., College Station, TX) and repeat it for every variable I want to test:

a. *Bootstrap "t test ppoFEV1, by(mortality)" r(P), reps(1,000) sav(name) replace*
   Where ppoFEV1 is the variable I want to test for the association with mortality and name is a name of your choice to save the bootstrap file. The t test can be replaced by any other test. For instance in case of non normal distribution of the numeric variable of interest a Wilcoxon rank-sum (or Mann Whitney test) can be used (command rank-sum instead of t test).
   The program saves the P values as _bs1_1
b. *Use name, clear*
c. *Count if _bs1_1<0.05*

This final command will return the number of times P value is less than 0.05 in 1,000 samples.

Internal validation of regression models

One of the most common use of bootstrap is regression model validation.

If applied to regression analysis, bootstrap can provide variables that have a high degree of reliability as independent risk factors (3).

We recently demonstrated that internal validation using resampling technique is superior to the traditional training and testing method (4). In this latter method, the original sample is randomly split in two sets, a development and a validation set. Most commonly 60% of the data at hand are used to develop the model and the other 40% to test the performance and validate the regression model.

This approach may be affected by selection bias and results may greatly vary owing to pure chance.

We compared the performance of a risk adjusted mortality model developed from the entire dataset of patients submitted to major lung resection and validated by bootstrap with that of ten different mortality models developed by using the traditional training-and-test method from the same dataset.

The performance of these eleven mortality models was tested by using the c-statistics in an external population of patients operated in another center.

Seventy percent of the models derived by the training and test method included different combinations of variables.

Their performances were extremely variable from one model to another, and in general, only modest compared to the model derived by using the entire sample validated by bootstrap. The latter method appears therefore much superior and reliable to develop reproducible and stable risk models to be applied in external populations.

One of the great advantages of using bootstrap to validate a regression model is that the entire original dataset can be utilized to develop more robust regression equations. This appears particularly important in moderate-size databases and for rare outcomes (i.e., mortality after major lung resection).

In practice, after a careful variables screening, a selected set of independent variables is entered in a regression analysis.

Then, a random sample of cases is selected with replacement, most commonly of the same size as the original sample. Regression is performed using this random simulated sample. This process is completely automated and the results of the analysis are stored. Then, another random sample of the same size is drawn from the original dataset with replacement (bootstrap) and regression is performed again in this new simulated sample of patients. This resampling of the original data

set followed by analysis continues most typically hundreds of times (typically 1,000 times). Finally, the frequency of occurrence of risk factors among these many models is summarized. Interestingly, each regression analysis performed in different simulated samples of observations usually generates different models containing different predictors. However, some predictors never turn out significant and others do so more frequently.

Usually predictors that result significant in more than 50% of bootstrap samples can be considered reliable and can be included in the final regression model.

The entire process is completely computer-automated and relies on specific statistical software and programs.

Bootstrap resampling allows to removing much of the human biases associated with regression analysis. It is a reliability test, which is able to eliminate or minimize the risks of selecting unreliable variables (type I error) and excluding reliable ones (type II error) (3).

To this purpose I use the following programming syntax in the Stata 12 statistical software (Stata Corp., College Station, TX):

Suppose I want to develop a risk model to identify risk factors associated with in-hospital mortality after lung resection. After a careful univariable screening of variables the following factors resulted associated with mortality: age, ppoFEV1, body mass index (BMI), pneumonectomy, predicted postoperative carbon monoxide lung diffusion capacity (ppoDLCO), induction chemotherapy.

I start with programming my logistic regression as follows:

a. *Program myreg, eclass*
b. *Logit mortality age ppoFEV1 BMI pneumonectomy ppoDLCO induction chemotherapy*
c. *Test age=0*
d. *Eret scalar Page=r(p)*
e. *Test ppoFEV1=0*
f. *Eret scalar ppoFEV1=r(p)*

……

q. end

Then I call myreg and verify the regression results. Subsequently I run the bootstrap command as follows:

a. *Bootstrap myreg _b Page=e(Page) ppoFEV1=e(ppoFEV1) ….., reps(1,000) sav(name)*
b. *Use name, clear*
c. *Count if Page<0.05*

This last command return the number of times age results significantly associated (P<0.05) with mortality in 1,000 bootstrap samples.

Model validation

Resampling with replacement can be also used to test a model on an external population not just once but hundreds of times.

One example is to verify performance of a risk score. In a recent paper, we developed an aggregate risk score to predict prolonged air leak (PAL) after lobectomy (5). We used a population of over 600 patients to develop the risk score, which was created by proportionally weighting the regression coefficients of the significant variables. According to their scores, patients were grouped in four risk classes with an incremental risk of developing PAL. The aggregate score was then validated in an external population of 230 patients operated on in another center. Instead of testing the score in one validation set, the latter one was bootstrapped to obtain 1,000 simulated external samples of patients of the same number of patients as the original validation set [230] and obviously different from the ones used to derive the model (derivation set). The frequency of occurrence of PAL in each class of risk was then tested in each of these 1,000 bootstrapped samples. For instance we found that in class A (the class with the lowest risk of PAL) 98% of samples had a risk of PAL less than 5%, whereas in class D (the highest class of risk) 99% of samples showed a PAL incidence of greater than 20%. This process showed that the score performed reliably across multiple populations and it is well suited to be used outside the set of patients from which it was derived.

Conclusions

Bootstrap resampling procedure is a computer intensive simulation technique that is capable to minimize much of the human arbitrariness from multivariable analysis and other analytical statistical techniques. It provides another important statistic: a measure of reliability of a risk factor that should complement the traditional calibration and discrimination measures of performance of a multivariable model. This measure should always be reported along with the magnitude of effect, its variance and P value in a regression table.

Acknowledgements

Disclosure: The author declares no conflict of interest.

References

1. Efron B. Bootstrap methods: another look at the jackknife. Ann Statist 1979;7:1-26.

2. Efron B, Tibshirani RJ. eds. An introduction to the bootstrap. New York: Chapman and Hall/CRC, 1993.
3. Blackstone EH. Breaking down barriers: helpful breakthrough statistical methods you need to understand better. J Thorac Cardiovasc Surg 2001;122:430-9.
4. Brunelli A, Rocco G. Internal validation of risk models in lung resection surgery: bootstrap versus training-and-test sampling. J Thorac Cardiovasc Surg 2006;131:1243-7.
5. Brunelli A, Varela G, Refai M, et al. A scoring system to predict the risk of prolonged air leak after lobectomy. Ann Thorac Surg 2010;90:204-9.

Cite this article as: Brunelli A. A synopsis of resampling techniques. J Thorac Dis 2014;6(12):1879-1882. doi: 10.3978/j.issn.2072-1439.2014.09.09

