Indian Journal of Fundamental and Applied Life Sciences ISSN: 2231– 6345 (Online)
An Open Access, Online International Journal Available at www.cibtech.org/sp.ed/jls/2015/01/jls.htm
2015 Vol.5 (S1), pp. 2414-2431/Sofla and Maleki
Research Article

USING BOOTSTRAPPING FOR EVALUATING THE PERFORMANCE OF DECISION MAKING UNITS BY DATA ENVELOPMENT ANALYSIS

Fatemeh Zalani Sofla and *Nader Rafati Maleki
Department of Mathematics, Tabriz Branch, Islamic Azad University, Tabriz, Iran
**Author for Correspondence*

ABSTRACT
Bootstrapping non-parametric models is a fairly complicated exercise which is associated with implicit assumptions or requirements that are not always obvious to the non-expert user. Bootstrap DEA is a significant development of the past decade; however, some of its assumptions and properties are still quite unclear, which may lead to mistakes in implementation and hypothesis testing. This paper clarifies these issues and proposes a hypothesis testing procedure, along with its limitations, which could be extended to test almost any hypothesis in bootstrap DEA. This article is based on a paper by DEA bootstrap and hypothesis testing of Simar and Wilson was drafted as a guide for users of DEA bootstrapping as a complementary activities Simar and Wilson, especially when the test is assumed to act as a deep understanding of Bootstrap DEA provides important functions to perform hypothesis testing problem has been dealt with using the bootstrap efficiency scores.

Keywords: *Bootstrap, Data Envelopment Analysis (DEA), Hypothesis Testing, DEA Bootstrap, Standard Error, Confidence Interval*

INTRODUCTION
Data Envelopment Analysis (DEA) is a non parametric method for calculating the size of performance of a group of units which have the same activity. This method has characteristics that cause this method to have some weak points. One of these weaknesses is the fragility of the efficiency obtained from this method. Changing the sample, may lead to a dramatic change in the obtained efficiency for units. Bootstrapping is a process which is used to overcome this problem. Estimating the amount of skew, skew-corrected estimates, confidence intervals and hypothesis testing, Boot Strapping is used to evaluate this uncertainty.
One of the most frequent uses of bootstrap DEA is to test various hypotheses. Since hypothesis test is a decision making tool based on observed sample related to the population that is criterion based on which sample decides about the population, therefore, it is not free from error. The most common errors found in the literature relate to: (i) the use of theoretically inconsistent hypothesis testing procedures, (ii) the use of potentially in appropriate tests in hypothesis testing, and (iii) applying regression analysis using the bootstrapped efficiency scores.
In this paper we provide these clarifications and we introduce a universal approach for using the bootstrapped efficiency scores in a theoretically consistent way, and comprise a valuable tool for implementing statistical inference on DEA.

Data Envelopment Analysis (DEA)
Data Envelopment Analysis (DEA) is a non parametric method for calculating the size of performance of a group of units which have the same activity. This method was first introduced by Charnes *et al.,* (1978) inspired by the work of Ferrier and Hirschberg (1997). They developed the proposed discussion by Farrell for the decision making units with multidimensional inputs and outputs in the absence of market prices. Farrell, under the hypothesis of Returns to Scale, considered the smallest convex hull which encompassed all sample data as the efficient boundary. The obtained boundary using this method is an accessible relative boundary in the real world. Unlike parametric method in which units are measured according to a boundary which is generally inaccessible in the real world, the efficiency obtained from this method was a relative amount not its real quantity. In other words, the efficiency obtained from this method was a projection of its real quantity. This boundary is built with the help of five principles including

observations, possibilities, endless- ray, convexity and minimum interpolation and it is called non parametric boundary of DEA. Those units which located on the boundary of DEA were efficient units and naturally other units were inefficient units.

The major advantage of DEA is that it does not require the specification of a production function: it just uses a set of inputs that DMUs want to minimize and a set of outputs that DMUs want to maximize.

Statistical Findings Related the Boundary of DEA

Banker (1993), considering the special function for efficient boundary showed the obtained boundary was the same stated boundary by DEA. He showed that the estimator of DEA had a weak adaptability and he also showed that maximum like hood estimator (MLE) was like the main boundary. Banker was the first person who presented statistical inferences on DEA boundary. Before presenting Baker's findings, DEA was criticized and attacked by many researchers due to being non parametric and failure in presenting statistical analysis. Although Banker achieved considerable results for DEA, he made no mention of DEA convergence speed to unknown boundary.

Korostelev *et al.,* (1995a and 1995b) considering Banker's results, investigated adaptability and convergence of DEA to a main unknown boundary in an output. Moreover, they showed that under weak general condition, DEA was Maximum Like hood Estimator (MLE) and it was based on Banker's findings. DEA boundary is the best estimator between convex borders with smooth bound and speed of it convergence $O(n^{-\frac{2}{p+2}})$.

Technical efficiency, as termed in DEA, is most commonly examined under the assumption of either input or output orientation. Under input orientation, DEA efficiency scores are interpreted as required input contractions to make a DMU efficient, keeping the level of outputs fixed. Under output orientation efficiency scores correspond to required output expansions to make a DMU efficient, keeping input levels fixed1. Hence, in input orientation inputs behave as variables and outputs as model parameters, while in output orientation outputs are the variables and inputs the constants. In this paper we will be using the CRS technology assumption under input orientation, although the extension to the output oriented case or VRS should be straightforward.

Bootstrap

One of the weaknesses of DEA is the fragility of the performance obtained by this method. One of these weaknesses is the fragility of the efficiency obtained from this method. Bootstrapping is a process which is used to overcome this problem (Bahar *et al.,* 2011)**.** In this section, we briefly introduce a bootstrapping process and the use of DEA models described above.

Bootstrap was first developed by Efron (1979) for estimating accuracy and distribution of sample statistics. Therefore, in regression plans, Bootstrap is usually preferred. Bootstrap idea is re-sampling of data. The bootstrap is a procedure of drawing with replacement from a sample, mimicking the data generating process of the underlying true model and producing multiple estimates which can be used for statistical inference. One of its most important uses is to test hypotheses, especially in cases where statistical inference is impossible otherwise. Resampling, within the framework of the bootstrap, relates to redistributing the assumed randomness of the model among observations. This randomness is reflected in the deviations of the model's variables from their expected values, as calculated (or estimated) by the model.

The Most Important Measures taken about Bootstrap

Bootstrap is a simulation method based on investigated data which is used for presenting statistical analysis. Efron (1993) chose the name" Bootstrap" because the last thing a soldier can utilize is a bootstrap. It refers to the story of Barron Van Manchmawsen who saved himself from a swamp with the help of his boot strap. Boot Strap has produced data in a complicated condition (such as non parametric models) without imposing any hypothesis and in a natural way. It considers the experimental distribution of data as the main and unknown distribution (Bahari *et al.,* 2011).

Bootstrap was introduced by Efron (1979) and developed by Efron and Tibshirani (1993). A few years after introducing Bootstrap, Leopold (1992) used it for estimating efficiency related to panel data in non

parametric models. The Salience of Bootstrap is related to Simar and Wilson (1998). They used Bootstrap for the first time to analysis the amount of efficiency sensitivity of DEA. They presented a smooth version of Bootstrap which removed the disadvantages of Bootstrap introduced by Efron (Nave Bootstrap) in border models. In addition in this article and in hat of Simar and Wilson (2000b) who developed the algorithm in Simar and Wilson (1998), they presented a general framework for estimating data generation process in non parametric models. It is worth mentioning that Ferrier and Hirschberg (1997) have used Bootstrap for estimating efficiency obtained from linear planning but their method was inadaptable and that of Simar-Wilson was the most comprehensive and completed one regarding boundary models.

Following works in this case generally include usages of Bootstrap or discussions on positive or negative factors affecting inadaptability of Bootstrap process and removing or enhancing them.

In some examples and real examples we have seen that some outputs are ineffective in the generation process, and existence and non existence of them have no effect on output amount. Also, we may want to know whether it is possible to gather some certain outputs and inputs together, considering the problems related to large dimensions of the problem. Nevertheless, it should be noted that before proceeding to sum up the input and output, they should be in the same scale. Simar-Wilson (2001) sought to test the accuracy of this subject by developing the hypothesis test.

Among other usages of Bootstrap conducted by Simar and Wilson (1999c), we can refer to the estimated productivity index of Malmquist and estimating of its analysis.

Communications Bootstrap and DEA

Although bootstrap DEA is not a recent development, some assumptions or requirements related to its implementation have not been clarified yet, while quite a few authors fail to use these methods appropriately.

The concept of efficiency has been traditionally related to the ratio of outputs over inputs of a certain firm relative to others. However, in a multiple input-output setup it is necessary to attach weights to inputs and outputs, which reflect their relative rate of usage, in order to calculate the ratio of weighted outputs over weighted inputs. DEA is a non-parametric technique which is based on this logic and uses linear programming to determine optimal weights which minimize the distance between the frontier and the decision making unit (DMU) under consideration, subject to disposability and convexity constraints.

One of the disadvantages of DEA is that statistical inference is very difficult to be applied on DEA scores. Therefore, bootstrap DEA was introduced by Simar and Wilson (1998), allowing to extract the sensitivity of efficiency scores which results from the distribution of (in) efficiency in the sample. Again, we would like to avoid demonstrating the technical details of the method since it is fairly established, while it would destruct the informed reader from the purpose of the paper. However, further details and analysis on related issues can be found in the papers of Simar and Wilson (1998, 1999, 2000a, 2000b) as well as their book chapters (Simar and Wilson, 2004, 2007, 2008). The outline of their proposed bootstrap procedure can be summarized in the following steps:

i. Use DEA to calculate efficiency scores.

ii. Draw with replacement from the empirical distribution (ED) of efficiency scores. Simar and Wilson (1998) suggest that smoothing the ED provides more consistent results.

iii. Divide the original efficient input levels by the pseudo-efficiency scores drawn from the (smoothed) empirical distribution to obtain a bootstrap set of pseudo-inputs.

iv. Apply DEA using the new set of pseudo-inputs and the same set of outputs and calculate the bootstrapped efficiency scores.

v. Repeat steps ii-iv B times and use bootstrapped scores for statistical inference and hypothesis testing.

The Logic behind Bootstrap DEA

The logic of bootstrapping within a model framework, applies to a large extent in the case of DEA. The choice between bootstrapping “pairs” (case resampling) or “residuals” (fixed resampling) depends on the model of DEA we are using. In oriented models, where either inputs or outputs are fixed, it is more reasonable to use fixed resampling, while in non-oriented models such as the additive model, it is more reasonable to apply case resampling.

Research Article

Statistics is changing. Modern computers and software make it possible to look at data graphically and numerically in ways previously inconceivable. They let us do more realistic, accurate, and informative analyses than can be done with pencil and paper. The bootstrap, permutation tests, and other resampling methods are part of this revolution. Resampling methods allow us to quantify uncertainty by calculating standard errors and confidence intervals and performing significance tests. They require fewer assumptions than traditional methods and generally give more accurate answers (sometimes very much more accurate). Moreover, resampling lets us tackle new inference settings easily. For example, for inference about the difference between two populations means. Resampling also helps us understand the concepts of statistical inference.

The sampling distribution is an abstract idea. The bootstrap analog (the "bootstrap distribution") is a concrete set of numbers that we analyze using familiar tools like histograms. Resampling methods for significance tests have the same advantage.

Bootstrap is a method which regardless of may hypotheses, makes the condition of the sample closer to the population by creating many samples and regarding all states of forming sample, one can make sure of the accuracy of the estimated co-efficient and estimating confidence intervals for co-efficient and estimating confidence intervals for co-efficient.(Efron and Tibshirani, 1993). When this method is used for abnormal data, it is of a great advantage (Henderson, 2005).

Suppose, we have an original sample of N, Bootstrap processing starts from this sample, a new random sample with the same size as the original sample is extracted (renewed sample) and meanwhile every selected observation after issuing returns to the original sample. This sampling is renewed and is the basis of Bootstrap. The bootstrap was introduced as a computer-based method for estimating the standard error of $\hat{\theta}$. It enjoys the advantage of being completely automatic. The bootstrap estimate of standard error requires no theoretical calculations. The bootstrap provides accuracy estimates by using the plug-in principle to estimate the standard error of a summary statistic. Bootstrap can be used to build statistical hypothesis test. This method is usually used as an alternative for inferential method based on parametric hypothesis at the time of any doubt about these hypotheses. Also, when calculating error standard becomes complicated, Bootstrap can be used. Significant tests are part of inferential statistics, which based on finding about sample, overgeneralization is made about the population. Hypothesis test refers to Sir Renal Fisher's studies (19^{th} and 20^{th} centuries, Jersey Neiman, 19^{th} century) and Karl Pearson (19^{th} and 20^{th} centuries). The modern hypothesis test is a combination of their works which are regarded as the 20^{th} century hypothesis test (Anders, 1998). In this paper we introduce a universal approach for using the bootstrapped efficiency scores in a theoretically consistent way. We focus our analysis on using bootstrap DEA to test the hypothesis of significant efficiency differences between two firms and we propose a straightforward and theoretically consistent procedure which can be easily extended to test any hypothesis.

Hypothesis Test

Any rule about population and population distribution or population parameter is called statistical hypothesis and may be true or false. A true or false hypothesis should be investigated based on information obtained from sampling of population and it is called hypothesis test.

Since claim can be true or false, two complementary hypotheses may be obtained: one for the claim to be true and the other for the claim to be false. Therefore, the starting of a hypothesis test should always include two statistical hypotheses which are located against each other.

In the discussion related to hypothesis test, we often face claims related to the distribution parameters of populations. These claims or hypotheses are called null hypothesis and it is shown by $H_\circ$. Statistical hypothesis which is against null hypothesis is called the opposite hypothesis and it is shown by H_1.

In general, hypothesis tests are shown as follows:

1) $\begin{cases} H_\circ : \theta = \theta_\circ \\ H_1 : \theta = \theta_1 \end{cases}$ 2) $\begin{cases} H_\circ : \theta = \theta_\circ \\ H_1 : \theta \neq \theta_0 \end{cases}$ 3) $\begin{cases} H_\circ : \theta \geq \theta_\circ \\ H_1 : \theta < \theta_1 \end{cases}$ 4) $\begin{cases} H_\circ : \theta \leq \theta_\circ \\ H_1 : \theta > \theta_1 \end{cases}$

Research Article

In the discussion related to statistical inference i.e., data analysis and over generalizing its results to the population of the study, there is also decision making regarding the proposed claims. In other words, research hypotheses which are proposed by appropriate methods (hypothesis tests) should be investigated and confirmed or rejected. In this case, the role of chance in data using probable patterns of sampling is emphasized.

Hypothesis test includes two parts of hypothesis or the proposed claim and test is a tool for investigating this claim which is confirmed or rejected based on research test.

Since hypothesis test is a tool for decision making based on observed sample regarding the population, i.e., it is a criterion based on which decision making is done, therefore, it is not free from errors.

In conducting a test, we finally make a decision to confirm or reject hypothesis $H_\circ$. We regard confirming $H_\circ$ as rejecting hypothesis H_1 and rejecting $H_\circ$ as confirming H_1. Based on decision regarding confirming or rejecting $H_\circ$, the following errors may occur:

Error type 1: rejecting hypothesis $H_\circ$ while hypothesis $H_\circ$ is true, is called error type one. The corresponding probability of error type one is shown by α and is defined as follow:

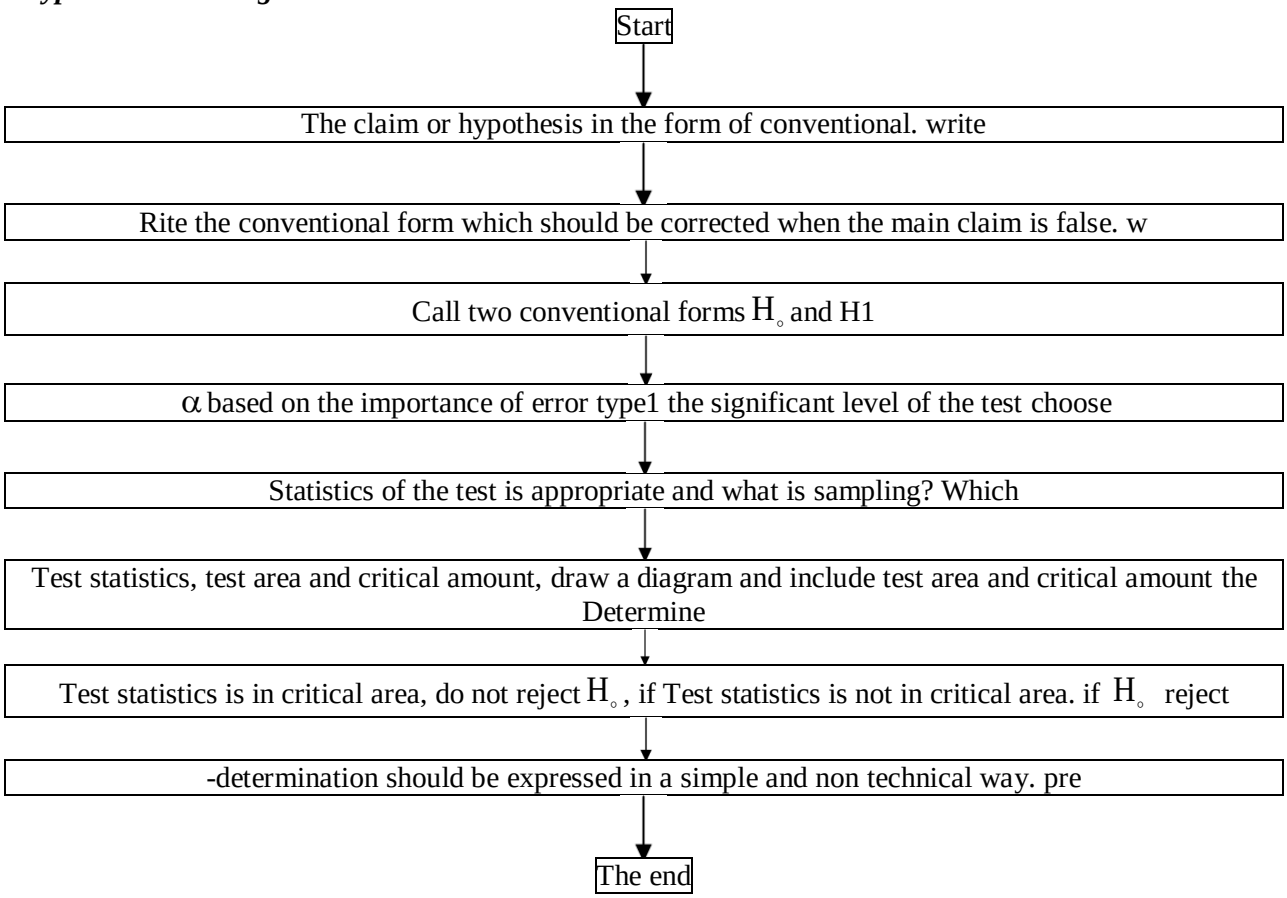
$$\alpha = P(errortype1) = P(H_\circ \ rejecj \mid true\, H_\circ)$$

Error type 1 is the significant level or recognition level. Confirming hypothesis $H_\circ$

While $H_\circ$ is not true is called error type2. The corresponding probability of error type2 is shown by β and is defined as follow:

$$\beta = P(errortype2) = P(H_\circ \ true \mid rejected \ H_\circ)$$

Hypothesis Test Stages



Hypothesis Testing with the Bootstrap

One of the most frequent uses of bootstrap DEA is to test various hypotheses. The literature on testing hypotheses using bootstrap DEA is either not very clear or limited to specific examples. Simar and Wilson (2008) provide guidance on using their techniques and demonstrate an example of hypothesis testing for the case of mean efficiency score differences between two groups. Among their general rules they suggest that: the test statistic used has to be a function of the data, the critical value should result from the bootstrap distribution while the null hypothesis and the alternative should be clearly stated and be theoretically sensible. However, it is not straightforward how one could use their methods to test hypotheses and which should be the principles which should be respected when testing hypotheses.

In this section we will provide an outline for designing and implementing hypothesis testing using the bootstrap distribution of efficiency scores.

Suppose that we want to test whether the DEA score of DMU A $(\hat{\theta}_A)$ differs significantly from the DEA score of DMU B $(\hat{\theta}_B)$, due to their sensitivity imposed by the distribution of (in) efficiency. Using the distribution of bootstrapped efficiency scores we could construct an acceptance region for DMU A and calculate the probability of observing the efficiency score of DMU B within this region. Hence, the hypothesis to be tested is:

$$H_\circ : \hat{\theta}_A = \hat{\theta}_B \quad , \quad H_1 : \hat{\theta}_A \neq \hat{\theta}_B \tag{1}$$

If the desired significance level is _ then we could use the $(\frac{a}{2})$ *and* $(1-\frac{a}{2})$ percentiles of the bootstrap distribution $(\hat{\theta}_A^b)$ in our two-tailed test. If we denote these percentiles with $\hat{p}_{\frac{a}{2}}$ *and* $\hat{p}_{1-\frac{a}{2}}$, respectively, we have:

$$pr\,(\hat{p}_{\frac{a}{2}} < \hat{\theta}_A^b < \hat{p}_{1-\frac{a}{2}}) = 1-a \tag{2}$$

Using straightforward manipulations (subtracting $\hat{\theta}_A^b$ and adding $\hat{\theta}_A$) we get:

$$pr\,(\hat{\theta}_A + \hat{\theta}_A^b - \hat{p}_{1-\frac{a}{2}} < \hat{\theta}_A^b < \hat{\theta}_A + \hat{\theta}_A^b - \hat{p}_{\frac{a}{2}}) = 1-a \tag{3}$$

Hence, we have constructed from (7) a $(1-a)\%$ region where the efficiency score of DMU A is expected to be observed, using the distribution of its bootstrapped efficiency scores. This implies that we could use a related p-value to calculate the probability of observing the DEA score of DMU B within the "region" of DMU A:

$$P = \frac{\#(\hat{\theta}_A + \hat{\theta}_A^b - \hat{p}_{1-\frac{a}{2}} < \hat{\theta}_B < \hat{\theta}_A + \hat{\theta}_A^b - \hat{p}_{\frac{a}{2}})}{B} \tag{4}$$

This is a standard indicator function used in bootstrap applications where the hash sign stands for "number of times". As usual, if $P > a$ the null hypothesis of no difference cannot be rejected. This straightforward logic can be extended to test any hypothesis.

One of the most troublesome limitations of this approach, which is common to all statistics or tests on bootstrap DEA, relates to the fact that the distribution of efficiency scores of each

DMU is not normal (in most cases skewed) and at the same time it is not identical to that of other DMUs even of the same sample. The importance of this result is that:

$$P_{AB} = \frac{\#(\hat{\theta}_A + \hat{\theta}_A^b - \hat{P}_{A,1-\frac{a}{2}} < \hat{\theta}_B < \hat{\theta}_A + \hat{\theta}_A^b - \hat{P}_{A,\frac{a}{2}})}{B}$$
$$\neq \frac{\#(\hat{\theta}_A + \hat{\theta}_A^b - \hat{P}_{B-1-\frac{a}{2}} < \hat{\theta}_A < \hat{\theta}_B + \hat{\theta}_B^b - \hat{P}_{B,\frac{a}{2}})}{B} = P_{BA} \quad (5)$$

Hence, the calculated p-value of (9) will differ depending on the reference DMU; that is, it is possible that $P_{BA} < a \ \ and \ \ P_{AB} > a$ and vice versa. In the case of skewed distributions it is preferable to use the median of the bootstrapped distribution while it is necessary to apply alternative methods to construct confidence intervals. Such methods have been proposed by

Efron (1982, 1987). However, although these methods improve the endpoints of the confidence intervals, it is still possible that the aforementioned problem persists. We therefore suggest for these few cases to reject the null hypothesis, since this seems to be a more conservative decision compared to accepting it.

Another limitation of this approach is that the extension to different samples requires the two samples to have similar distribution of inefficiency (to ensure similar source of variability).

To some extent we could mitigate this issue by applying our test on the standardized efficiency scores, although the higher moments (skewness and kurtosis) would still need to be similar. Hence, if the standardized efficiency score of any DMU $\mathrm{K} = \{\mathrm{A}, \mathrm{B}\}$ is $\hat{\varsigma}_{\mathrm{k}} = \frac{(\hat{\theta}_{\mathrm{k}} - \overline{\hat{\theta}^{(\mathrm{k})}})}{\mathrm{S}^{(\mathrm{k})}(\hat{\theta})}$, where $\overline{\hat{\theta}^{(k)}}$ *and* $S^{(k)}(\hat{\theta})$ are the mean and the standard deviation of the efficiency scores of the group where DMU k belongs to, then (8) becomes:

$$Pr\left(\frac{\hat{\theta}_A + \hat{\theta}_A^b - \hat{P}_{1-\frac{a}{2}} - \overline{\hat{\theta}^{(A)}}}{S^{(A)}(\hat{\theta})} < \hat{\varsigma}_A < \frac{\hat{\theta}_A + \hat{\theta}_A^b - \hat{P}_{\frac{a}{2}} - \overline{\hat{\theta}^{A}}}{S^{(A)}(\hat{\theta})}\right) \quad (6)$$

Then we could substitute for $\hat{\varsigma}_{\mathrm{B}}$ and calculate the associated p-value as in (9), which should be exactly the same if DMUs A and B were from the same sample. However, if they do not, the two different groups need to be comparable (ideally homogeneous) while any differences in their means should be assumed to be random. Hence, standardizing would ensure that both groups have the same mean (zero) and variance (one), while the resulting variables would be comparable as they reflect standardized deviations from the mean. Note, though, that the skewness and kurtosis of the standardized efficiency scores are identical to the non standardized ones, which supports our previous argument that higher moments still need to be similar to get meaningful results.

Standard Errors and Estimated Standard Errors

Summary statistics such as $\hat{\theta} = t\left(\hat{F}\right)$ are often the first outputs of a data analysis. The next thing we want to know is the accuracy of $\hat{\theta}$. The bootstrap provides accuracy estimates by using the plug-in principle to estimate the standard error of a summary statistic. First we will discuss estimation of the standard error of a mean, where the plug-in principle can be carried out explicitly.

Suppose that x is a real-valued random variable with probability distribution F. let us denote the expectation and variance of F by the symbols $\mu_F \ and \ \sigma_F^2$ respectively,

$$\mu_F = E_F(x)\ ,\ \sigma_F^2 = \mathrm{var}_F(x) = E_F[x-\mu_F)^2]. \tag{7}$$

The alternative notation "$\mathrm{var}_F(x)$" for the variance, sometimes abbreviated to *var(x)*, means the same thing as σ_F^2. In what follows we will sometimes write

$$x \sim (\mu_F\ ,\ \sigma_F^2) \tag{8}$$

To indicate concisely the expectation and variance of *x*. Now let $(x_1, x_2, ..., x_n)$ be a random sample of size *n* from the distribution F. The mean of the sample $\bar{x} = \sum_{i=1}^{n} x_i/n$ has expectation μ_F and variance

$$\sigma_F^2/n,\ \bar{x} \sim (\mu_F\ ,\ \sigma_{F/n}^2). \tag{9}$$

In other words, the expectation of $\bar{x}$ is the same as the expectation of a single *x*, but the variance of $\bar{x}$ is $1/n$ times the variance of *x*. This is the reason for taking averages; the larger *n* is, the smaller $\mathrm{var}(\bar{x})$ is, so bigger *n* means a better estimate of μ_F.

The standard error of the mean $\bar{x}$, written $se_F(\bar{x})$ or $se(\bar{x})$, is the square root of the variance of $\bar{x}$,

$$se_F(\bar{x}) = [\mathrm{var}_F(\bar{x})]^{\frac{1}{2}} = \frac{\sigma_F}{\sqrt{n}}. \tag{10}$$

Standard error is a general term for the standard deviation of a summary statistic. They are the most common way of indicating statistical accuracy. Under quite general conditions on F, the distribution of $\bar{x}$ will be approximately normal as *n* gets large, which we can write as

$$\bar{x} \sim N(\mu_F\ ,\ \sigma_{F/n}^2). \tag{11}$$

The expectation μ_F and σ_F^2/n variance in (11) are exact.

The Bootstrap Estimation of Standard Error

A random sample $(x_1, x_2, ..., x_n)$ from an unknown probability distribution F has been observed and we wish to estimate a parameter of interest $\theta = t(F)$ on the basis of *x*. For this purpose, we calculate an estimate $\hat{\theta} = s(x)$ from *x*.

Bootstrap methods depend on the notion of a bootstrap sample. Let $\hat{F}$ be the empirical distribution, putting probability $1/n$ on each of the observed values $x_i, i = 1, 2, ..., n$. A bootstrap sample is defined to be a random sample of size *n* drawn from $\hat{F}$, say $X^*(x_1^*, x_2^*, ..., x_n^*)$,

$$F \rightarrow (x_1^*, x_2^*, ..., x_n^*). \tag{12}$$

The star notation indicates that x^* is not the actual data set *x*, but rather a randomized, or resampled, version of *x*.

There is another way to say (12): the bootstrap data points $x_1^*, x_2^*, ..., x_n^*$ are a random sample of size *n* drawn with replacement from the population of *n* objects $(x_1, x_2, ..., x_n)$. Thus we might have $x_1^* = x_7, x_2^* = x_3, x_3^* = x_3, x_4^* = x_{22}, ..., x_n^* = x_7$.

The bootstrap data set $(x_1^*, x_2^*, ..., x_n^*)$ consists of members of the original data set $(x_1, x_2, ..., x_n)$, some appearing zero times, some appearing once, some appearing twice, etc.

Corresponding to a bootstrap data set x^* is a bootstrap replication of $\hat{\theta}$,

$$\hat{\theta}^* = s(X^*). \tag{13}$$

The quantity $s(X^*)$ is the result of applying the same function s(.) to x^* as was applied to x. For example if $s(X)$ is the sample mean $\overline{X}$ then $S(X^*)$ is the mean of the bootstrap data set, $\bar{x}^* = \sum_{i=1}^{n} \frac{x_i^*}{n}$.

The bootstrap estimate of $se_F(\hat{\theta})$, the standard error of a statistic $\hat{\theta}$, is a plug-in estimate that uses distribution function $\hat{F}$ in place of the unknown distribution F. Specifically, the bootstrap estimate of $se_F(\hat{\theta})$ is defined by

$$se_{\hat{F}}(\hat{\theta}^*) \tag{14}$$

In other words, the bootstrap estimates of $se_F(\hat{\theta})$ is the standard error of $\hat{\theta}$ for data sets of size *n* randomly sampled from $\hat{F}$.

Formula (14) is called the ideal bootstrap estimate of standard error of $\hat{\theta}$. Unfortunately, for virtually any estimate $\hat{\theta}$ other than the mean, there is no neat formula that enables us to compute the numerical value of the ideal estimate exactly. The bootstrap algorithm, described next, is a computational way of obtaining a good approximation to the numerical value of $se_{\hat{F}}(\hat{\theta}^*)$.

It is easy to implement bootstrap sampling on the computer. A random number device selects integers $i_1, i_2, ..., i_n$, each of which equals any value between 1 and *n* with probability $\frac{1}{n}$. The bootstrap sample consists of the corresponding members of *x*,

$$x_1^* = x_{i_1},\ x_2^* = x_{i_2}, \ldots,\ x_n^* = x_{i_n} \tag{15}$$

The bootstrap algorithm works by drawing many independent bootstrap samples, evaluating the corresponding bootstrap replications, and estimating the standard error of $\hat{\theta}$ by the empirical standard deviation of the replications. The result is called the bootstrap estimate of standard error, denoted by $\hat{se}_B$, where *B* is the number of bootstrap samples used.

The following algorithm is a more explicit description of the bootstrap procedure for estimating the standard error of $\hat{\theta} = S(X)$ from the observed data *x*.

The Bootstrap Algorithm for Estimating Standard Error

1. Select *B* independent bootstrap samples $x^{*1}, x^{*2}, ..., x^{*B}$, each consisting of *n* data values drawn with replacement from *x*, as in (12) or (15). [For estimating a standard error, the number *B* will ordinarily be in the ranges 25-200].
2. Evaluate the bootstrap replication corresponding to each bootstrap sample,

$$\hat{\theta}^*(b) = S(X^{*b}) \qquad b = 1, 2, ..., B \tag{16}$$

3. Estimate the standard error $se_F(\hat{\theta})$ by the sample standard deviation of the *B* replications

$$\hat{se}_B = \{\sum_{b=1}^{B} [\hat{\theta}^*(b) - \hat{\theta}^*(\circ)]^2 / (B-1)\}^{\frac{1}{2}} \quad , \quad \hat{\theta}^*(\circ) = \sum_{b=1}^{B} \hat{\theta}^*(b)/B \tag{17}$$

Figure 1 is a schematic diagram of the bootstrap standard error algorithm. The limit of $\hat{se}_B$ as *B* goes to infinity is the ideal bootstrap estimate of $se_F(\hat{\theta})$,

$$\lim_{B\to\infty} \hat{se}_B = se_{\hat{F}} = se_{\hat{F}}(\hat{\theta}^*) \tag{18}$$

The fact that $\hat{se}_B$ approaches $se_{\hat{F}}$ as *B* goes to infinity amounts to saying that an empirical standard deviation approaches the population standard deviation as the number of replications grows large. The 'population' in this case is the population of values $\hat{\theta}^* = S(X^*)$, where $F \to (x_1^*, x_2^*, \ldots, x_n^*) = X^*$.

The ideal bootstrap estimate $se_{\hat{F}}(\hat{\theta}^*)$ and its approximation $\hat{se}_B$ are sometimes called nonparametric bootstrap estimates because they are base on $\hat{F}$, the nonparametric estimate of the population *F*.

If $s(X)$ is the sample median, for instance, then $s(X^*)$ is the median of the bootstrap sample. The bootstrap estimate of standard error is the standard deviation of the bootstrap replications,

$$\hat{se}_{boot} = \{\sum_{b=1}^{B} [s(X^{*b}) - s(\circ)]^2 / (B-1)\}^{\frac{1}{2}} \quad , \quad s(\circ) = \sum_{b=1}^{B} s(X^{*b}) / B \tag{19}$$

Confidence Interval Construction

In the previous sections we established that testing for efficiency score differences between

Two DMUs of the same sample is associated with the probability statement in (3):

$$pr(\hat{\theta}_A + \hat{\theta}_A^b - \hat{p}_{1-a/2} < \hat{\theta}_A < \hat{\theta}_A + \hat{\theta}_A^b - \hat{p}_{a/2}) = 1 - a \tag{20}$$

This information can be used to construct confidence intervals or acceptance regions about $\hat{\theta}_A$. Hence, if the efficiency score of another DMU falls within the region of DMU A we could state that the two DMUs do not differ significantly in efficiency and this will be due to the implied sensitivity of efficiency scores introduced by the distribution of (in) efficiency. To perform this task we will need to calculate two percentiles: one for the lower bound and one for the upper bound in (20). Denote the $(a/2)^{th}$ percentile of $\hat{\theta}_A + \hat{\theta}_A^b + \hat{p}_{1-a/2}$ with $\hat{t}_l$ and the $(1-a/2)^{th}$ percentile of $\hat{\theta}_A + \hat{\theta}_A^b + \hat{p}_{a/2}$ with $\hat{t}_u$. These percentiles are associated with the following one-tailed probability statements which we will need to use to construct our central $(1-a)\%$ confidence interval (Efron, 1982):

$$pr(\hat{t}_L < \hat{\theta}_A) = pr(\hat{\theta}_A < \hat{t}_u) = 1 - \frac{a}{2} \tag{21}$$

It is straightforward to verify that:

$$pr(\hat{t}_l < \hat{\theta}_A < \hat{t}_u) = 1 - a \tag{22}$$

Therefore the lower and upper bounds of our confidence intervals are $\hat{t}_u, \hat{t}_l$, respectively.

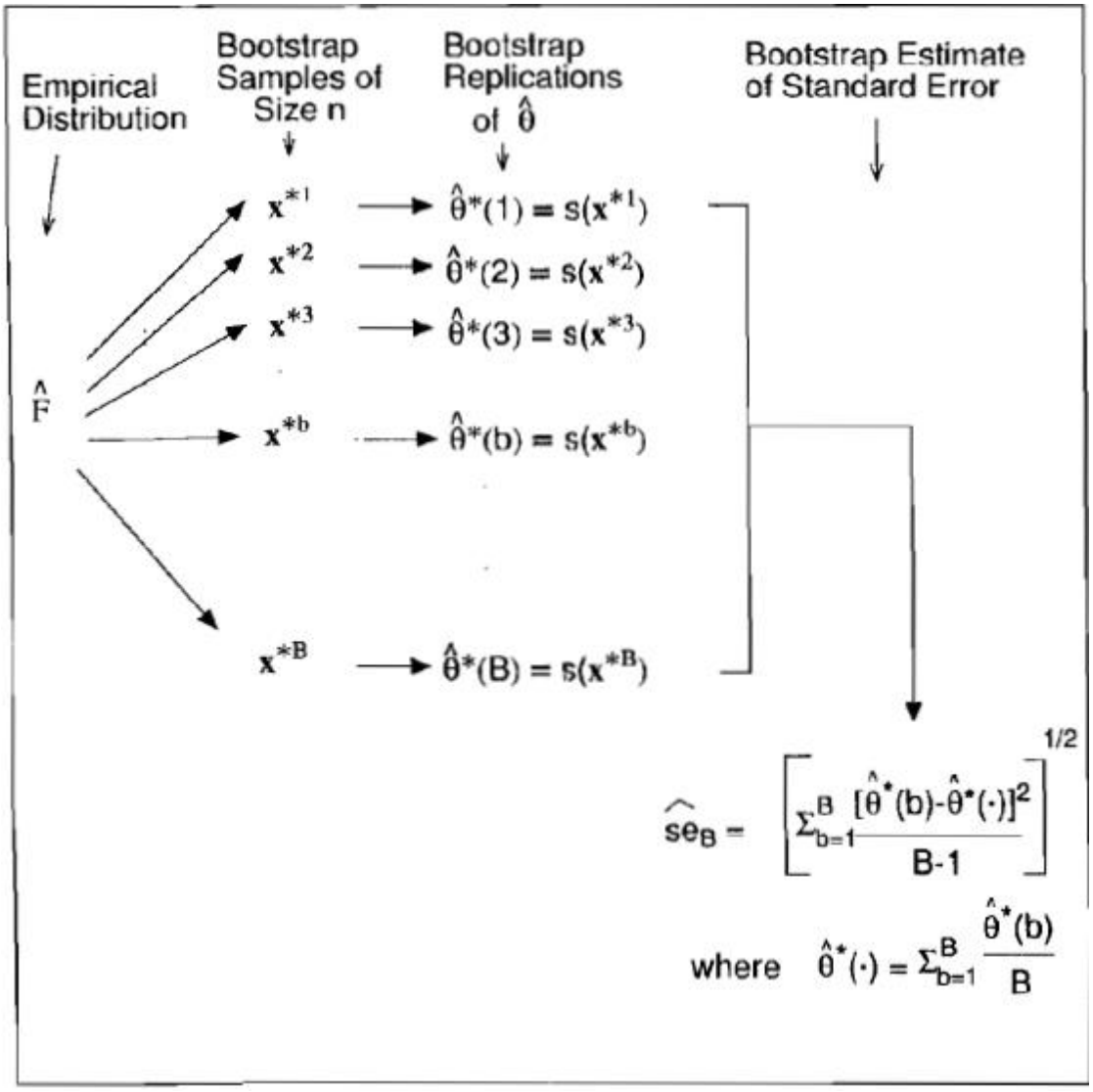


Figure 1: The bootstrap algorithm for estimating the standard error of a statistic $\hat{\theta} = s(X)$; each bootstrap sample is an independent random sample of size *n* from $\hat{F}$. The number of bootstrap replications *B* for estimating a standard error is usually between 25 and 200. As $B \to \infty$, $\hat{se}_B$ approaches the plug-in estimate of $se_{\hat{F}}(\hat{\theta})$

The confidence intervals will be centered on $\hat{\theta}_A$ by construction, if the medians of the Bootstrapped distributions are used in all relevant calculations (for example, for the bias). However, the bootstrap distributions of efficiency scores are usually skewed and the calculated confidence intervals will be biased to some extent. Therefore appropriate techniques should be implemented which correct for skewness and provide more accurate endpoints for the constructed confidence intervals. Simar and Wilson (1998) suggest using the bias corrected (BC) intervals of Efron (1982), however it is not the best option when dealing with skewness. A more appropriate method is that of Efron (1987), where the "bias corrected and accelerated" (BC_a) confidence intervals account for skewness through the *acceleration parameter*. The first step to construct the central BC_a confidence intervals with coverage $1-a$ is to calculate corrected

percentiles of the bootstrap distribution endpoints. Without loss of generality, if we are using our definition of the bias-corrected bootstrap distribution, that is $\hat{\theta}_A^{b*} = \hat{\theta}_A^b - b\hat{i}as_A$, we would be replacing the percentiles $\hat{S}_{\alpha/2}$ and $\hat{S}_{1-\alpha/2}$ with the BC_a ones $\hat{S}^{(a_1)}$ and $\hat{S}^{(a_2)}$, where

$$S^{(a_1)} = \phi\left(\hat{Z}_o + \frac{\hat{Z}_o + Z^{(\alpha/2)}}{1+\hat{a}(\hat{z}_o + z^{(\alpha/2)})}\right) \tag{23}$$

and

$$S^{(a_2)} = \phi\left(\hat{Z}_o + \frac{\hat{Z}_o + Z^{(1-\alpha/2)}}{1-\hat{a}(\hat{Z}_o + Z^{(1-\alpha/2)})}\right) \tag{24}$$

where ϕ is the standard normal cumulative density function and $Z^{(\alpha/2)}$ is the normalized value that corresponds to the $\alpha/2^{th}$ percentile of the standard normal distribution, so that $\phi(Z^{(\alpha/2)}) = \alpha/2$. The parameter $\hat{Z}_o$ is called the bias correction parameter and depends on the proportion of bootstrap estimates that are lower than the model estimates: $G(\hat{\theta}_A) = pr(\hat{\theta}_A^{b*} < \hat{\theta}_A)$ and $\hat{Z}_o = \phi^{-1}[G(\hat{\theta}_A)]$ is the standard normal value that corresponds to that probability. In our particular example $\hat{\theta}_A^{b*}$ is already bias-corrected by the median of its distribution, hence $G(\hat{\theta}_A^{b*}) = 0/5$ and therefore $\hat{Z}_o = o$.

To summarize, in order to obtain appropriate confidence intervals for the DEA score of DMU A, we suggest using the BC_a method of Efron (1987) to appropriately compute the endpoints a_1 and a_2 of the distribution of $\hat{\theta}_A^{b*}$ that is $\hat{\theta}_A^{b*,(a_1)}$, $\hat{\theta}_A^{b*(a_2)}$. Denote, then, the percentiles of this distribution as $\hat{S}^{(a_2)}, \hat{S}^{(a_1)}$ and by applying appropriate transformations we have:

$$\begin{aligned} &pr(\hat{\theta}_A + \hat{\theta}_A^b - \hat{p}^{(a_2)} < \hat{\theta}_A < \hat{\theta}_A + \hat{\theta}_A^b - \hat{p}^{(a_1)}) \\ &= pr(\hat{\theta}_A + \hat{\theta}_A^{b*} - \hat{S}^{(a_2)} < \hat{\theta}_A < \hat{\theta}_A + \hat{\theta}_A^{b*} - \hat{S}^{(a_1)}) = 1-a \end{aligned} \tag{25}$$

Like previously, we will need to use the a_1^{th} and a_2^{th} percentiles of the two endpoints in (25), which we denote as $\hat{t}^{(a_1)}$ and $\hat{t}^{(a_2)}$, respectively. Finally, the central BC_a percentiles with coverage $1-a$ are calculated by $\left(\hat{t}^{(a_1)}, \hat{t}^{(a_2)}\right)$.

To maximize intuition we have graphically represented in Figure 2 what bootstrap DEA does and how hypothesis testing is performed. In our simple one input (*x*), one output (*y*) case, we consider a sample of 30 DMUs which is randomly drawn from an underlying population. The sample CRS frontier is defined by DMU A, while the hypothesized (unobserved) population frontier is also drawn for comparison. Sampling bias in this case is considered to be the distance between the population frontier and the sample frontier. The bias is common to all DMUs and it is reflected on the fixed angle between the frontiers, the tangent of which reflects technical efficiency. The widening gap reflects the fact that the fixed efficiency score differential is translated into bigger input contractions as input levels increase, which is very reasonable.

If we focus our analysis on DMU A, which is assumed to be the most efficient DMU in the sample, bootstrapping its efficiency scores can be translated into varying its input levels. We can center this variation about $\hat{\theta}_A$ by correcting for bias $\left(\hat{\theta}_A^{b*}\right)$ and the resulting input variation is represented here by the horizontal dotted line. Furthermore, following the aforementioned procedure we may construct confidence intervals, to see which firms do not differ significantly in performance. This is represented by

the shaded area that is defined between the lower and upper bounds of the confidence interval. Note that we have taken care to intersect the horizontal dotted line close to the edges, leaving out some information at the tails. In our example we observe that about 3 DMUs fall within the confidence interval region, hence these DMUs do not differ significantly in efficiency from DMU A.

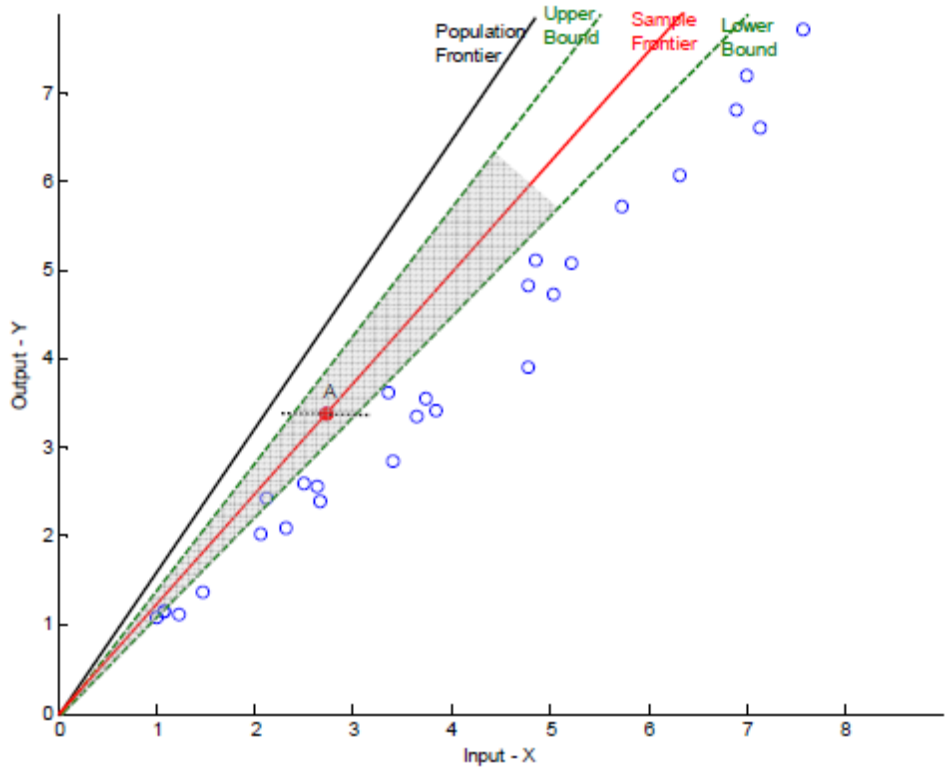


Figure 2: Graphical representation of hypothesis testing in bootstrap DEA

Bootstrap Confidence Intervals

There are several ways to construct bootstrap confidence intervals. Here we discuss three methods.

Method 1: The Normal Interval. The simplest method is the Normal interval

$$T_n \pm Z_{\alpha/2}\, se\, boot \tag{26}$$

Where $se_{boot} = \sqrt{v_{boot}}$ is the bootstrap estimate of the standard error. This Interval is not accurate unless the distribution of T_n is close to Normal.

Method 2: Pivotal Intervals. Let $\hat{\theta}_n = T(\hat{F}_n)$, $\theta = T(F)$ and define the **pivot** $R_n = \hat{\theta}_n - \theta$. *Let* $\hat{\theta}^*_{n,1},.....,\hat{\theta}^*_{n,B}$ denote bootstrap replications of $\hat{\theta}_n$. Let *H*(*r*) denote the CDF of the pivot:

$$H(r) = P_F(R_n \le r). \tag{27}$$

Define $C^*_n = (a,b)$ where

$$a = \hat{\theta}_n - H^{-1}(1-\alpha/2)\ ,\quad b = \hat{\theta}_n - H^{-1}(\alpha/2). \tag{28}$$

It follows that

$$\begin{aligned} P(a \le \theta \le b) &= p(a-\hat{\theta}_n \le \theta-\hat{\theta}_n \le b-\hat{\theta}_n) \\ &= p(\hat{\theta}_n-b \le \hat{\theta}_n-\theta \le \hat{\theta}_n-a) \\ &= p(\hat{\theta}_n-b \le R_n \le \hat{\theta}_n-a) \\ &= H(\hat{\theta}_n-a)-H(\hat{\theta}_n-b) \\ &= 1-\alpha/2-\alpha/2 = 1-a \end{aligned}$$

Hence, C_n^* is an exact 1 – *a* confidence interval for *θ*. unfortunately, *a* and *b* values depend on the unknown distribution *H* but we can form a bootstrap estimate of *H*:

$$\hat{H}(r) = \frac{1}{B}\sum_{b=1}^{B} I(R_{n,b}^* \le r) \qquad (29)$$

Where $R_{n,b}^* = \hat{\theta}_{n,b}^* - \hat{\theta}_n$. Let r_β^* denote the β sample quintile of $(R_{n,1}^*,\ldots,R_{n,B}^*)$ and let θ_β^* denote the β sample quantile of $(\hat{\theta}_{n,1}^*,\ldots,\hat{\theta}_{n,B}^*)$. Note that $r_\beta^* = \theta_\beta^* - \hat{\theta}_n$. It follows that an approximate $1-a$ confidence interval is $C_n = (\hat{a},\hat{b})$ where

$$\hat{a} = \hat{\theta}_n - \hat{H}^{-1}(1-\alpha/2) = \hat{\theta}_n - r_{1-\alpha/2}^* = 2\hat{\theta}_n - \theta_{1-\alpha/2}^*$$

$$\hat{b} = \hat{\theta}_n - \hat{H}^{-1}(\alpha/2) = \hat{\theta}_n - r_{\alpha/2}^* = 2\hat{\theta}_n - \theta^*{}_{\alpha/2}$$

In summary, the 1 – *a* **bootstrap pivotal confidence** interval is

$$C_n = (2\hat{\theta}_n - \hat{\theta}_{1-\alpha/2}^*,\ 2\hat{\theta}_n - \hat{\theta}_{\alpha/2}^*) \qquad (30)$$

Method 3: Percentile Intervals. The **bootstrap percentile interval** is defined by

$$C_n = (\theta_{\alpha/2}^*\ ,\ \theta_{1-\alpha/2}^*)$$

The justification for this interval is given in the appendix.

Sensitivity Analysis of the Original Efficiency Scores

In the usual application, the researcher is confronted with a set of observations $X = \{(x_i, y_i) \mid i = 1,\ldots,n\}$ corresponding to *n* production units. For each of the *n* observed units, we wish to analyze the sensitivity of the efficiency scores estimated by $\hat{\theta}_n \ldots, \hat{\theta}_1$.The procedure in previous section may be followed by allowing each observation $(x_k, y_k) : K = 1,\ldots,n$ to replace (X_o, Y_o) sequentially. This allows us to analyze the sensitivity of the distance from a fixed point (x_k, y_k) to the estimated frontier $\partial\widehat{X(y_k)}$, relative to the sampling variation of the estimator of the frontier, taking into account the entire set of observations X. For the DEA approach, the complete bootstrap algorithm is summarized by the following steps:

(1) For each $\hat{\theta}_k$, (x_k, y_k) ; $k = 1,\ldots,n$ compute $\hat{\theta}_k$ by the BCC linear program.

(2) Using the smooth bootstrap of previous section, generate a random sample of size *n* from $\hat{\theta}_i : i = 1,\ldots,n$ providing $\theta^*{}_{1b},\ldots,\theta_{nb}^*$.

(3) Compute $X_b^* = \{(x_{ib}, y_i); i = 1,\ldots,n\}$, where $i = 1,\ldots,n$, $X_{ib}^* = \left(\frac{\hat{\theta}_i}{\theta_{ib}^*}\right) X_i$.

(4) Compute the bootstrap estimate $\hat{\theta}_{k,b}^*$ of $\hat{\theta}_k$ for $K = 1,\ldots,n$ by solving

$$\hat{\theta}^{*}_{k,b} = Min\left\{\begin{array}{l} \theta \mid y_K \le \sum_{i=1}^{n}\delta_i y_i \quad \& \quad \theta X_K \ge \sum_{i=1}^{n}\delta_i \; x^{*}_{k,b} \;\; ;\theta > 0; \\ \sum_{i=1}^{n}\delta_i = 1 \; ; \; \delta_i \ge 0, i = 1,\ldots,n \end{array}\right\}$$

(5) Repeat steps 2–4 *B* times to provide for $k = 1,\ldots,n$ a set of estimates

$\left\{\hat{\theta}^{*}_{k,b} \;, b = 1,\ldots,B\right\}$.

For large datasets, the choice of *B* will be constrained by available computer resources. Hall (1986) suggests setting *B=1000* to ensure adequate coverage of the confidence intervals.

Empirical Illustration

To illustrate the methodology proposed in §11, we use data from Fare and Grosskopf (1989) on 19 electric utilities operating. The data contain information on one output (electric power, measured in KWh) and three inputs (labor, measured by average annual employment; fuel; and capital, represented by installed capacity measured in MW).

Table 1 shows the results for the bootstrap exercise for *B=1000* and h=0.014. Column 1 indicates the firm number, while columns 2–6 give the original DEA efficiency estimate, the bias-corrected estimate, the bootstrap bias estimate, the median of the bootstrapped values, and their standard deviation, respectively. The last four columns provide 95% confidence intervals for the bias-corrected efficiency estimates. The first confidence interval is based on the bias-correction formula in $\left(\hat{\theta}_{k,low}, \hat{\theta}_{k,up}\right) = \left(\tilde{\theta}^{*(a)}_{k}, \tilde{\theta}^{*(1-a)}_{k}\right)$, while the second confidence interval was computed from the median-centering device represented in $\left(\hat{\theta}_{k,low}, \hat{\theta}_{k,up}\right) = \left(\tilde{\theta}^{*(a_1)}_{k}, \tilde{\theta}^{*(a_2)}_{k}\right)$. Since in each case the median of $\tilde{\theta}^{*}_{k,b}$ is close to $\tilde{\theta}_k$, the two sets of confidence intervals are similar.

The results in Table 1 reveal the sensitivity of the efficiency measures w.r.t. sampling variation. The results indicate that one should be careful in making relative comparisons of the performances among firms based on the original DEA efficiency scores $\tilde{\theta}_k$. For example, Firm 1 has a DEA efficiency score $\hat{\theta}_1$ = 0.8692, while Firm 2 is ostensibly efficient with $\hat{\theta}$=1.0. With the bias-corrected measure in column 3, the difference is less dramatic, but still substantial.

However, the last four columns show that the confidence intervals for the efficiency of the two firms overlap to a large degree. Thus, we would not say that the two firms are significantly different in terms of their technical efficiency.

The two sets of confidence intervals based on $\left(\hat{\theta}_{k,low}, \hat{\theta}_{k,up}\right) = \left(\tilde{\theta}^{*(a)}_{k}, \tilde{\theta}^{*(1-a)}_{k}\right)$ and $\left(\hat{\theta}_{k,low}, \hat{\theta}_{k,up}\right) = \left(\tilde{\theta}^{*(a_1)}_{k}, \tilde{\theta}^{*(a_2)}_{k}\right)$ are very similar, with the median-centered intervals in the last two columns of Table 1 shifted slightly to the right relative to the mean centered intervals in columns 7–8.

While use of mean centering is probably more common in other bootstrap settings, the median provides a more robust measure of location than the mean when distributions are skewed as with DEA efficiency scores. The present example, however, indicates little practical difference in the two approaches.

Table 1: Bootstrap with bandwidth h=0.014

						2.5%	97.5%	2.5%	97.5%
k	$\hat{\theta}_k$	$\tilde{\theta}_k$	bias_k	Median of $\tilde{\theta}^*_{k,b}$	Std. Dev.	Bias Corrected		Centered on $\tilde{\theta}_k$	
1	0.8692	0.8519	0.0173	0.8480	0.0143	0.8360	0.8854	0.8372	0.9057
2	1.0000	0.9307	0.0693	0.9145	0.0614	0.8631	1.0564	0.8646	1.0800
3	1.0000	0.9457	0.0543	0.9396	0.0475	0.8932	1.0574	0.8937	1.0647
4	0.9307	0.9173	0.0133	0.9148	0.0094	0.9059	0.9400	0.9076	0.9536
5	1.0000	0.9349	0.0651	0.9177	0.0573	0.8717	1.0573	0.8730	1.0807
6	0.9071	0.8907	0.0165	0.8849	0.0151	0.8756	0.9286	0.8777	0.9552
7	0.8915	0.8759	0.0156	0.8693	0.0185	0.8615	0.9353	0.8652	0.9861
8	0.8210	0.8076	0.0135	0.8041	0.0118	0.7955	0.8387	0.7967	0.8630
9	0.8892	0.8624	0.0268	0.8488	0.0301	0.8370	0.9434	0.8405	0.9816
10	0.8469	0.8374	0.0095	0.8359	0.0064	0.8294	0.8541	0.8304	0.8598
11	0.9534	0.9423	0.0111	0.9396	0.0104	0.9325	0.9720	0.9339	1.0010
12	1.0000	0.9335	0.0665	0.9153	0.0591	0.8689	1.0484	0.8700	1.0749
13	0.9602	0.9434	0.0168	0.9389	0.0140	0.9282	0.9764	0.9304	1.0049
14	1.0000	0.9258	0.0742	0.9036	0.0676	0.8534	1.0786	0.8545	1.0872
15	1.0000	0.9334	0.0666	0.9085	0.0640	0.8683	1.0786	0.8697	1.1029
16	0.8885	0.8767	0.0117	0.8742	0.0091	0.8664	0.9022	0.8684	0.9169
17	1.0000	0.9378	0.0622	0.9179	0.0561	0.8774	1.0608	0.8783	1.0774
18	1.0000	0.9424	0.0576	0.9325	0.0499	0.8866	1.0527	0.8871	1.0646
19	0.9441	0.9327	0.0113	0.9305	0.0083	0.9228	0.9546	0.9238	0.9645

Tables 2 and 3 present similar results obtained with different values of the bandwidth h; in Table 2, the bandwidth is reduced by half, while in Table 3 the band width is doubled relative to the value used in Table 1. The results do not appear very sensitive with respect to the different bandwidths, although for h=0.007, more weight is given near the upper bound of θ, while for h=0.028, the distributions are shifted slightly to the left. This is reassuring, since the literature on kernel estimation presents a variety of objective functions that could be optimized to choose the bandwidth h.

Table 2: Bootstrap with bandwidth h=0.007

						2.5%	97.5%	2.5%	97.5%
k	$\hat{\theta}_k$	$\tilde{\theta}_k$	bias_k	Median of $\tilde{\theta}^*_{k,b}$	Std. Dev.	Bias Corrected		Centered on $\tilde{\theta}_k$	
1	0.8692	0.8560	0.0132	0.8498	0.0142	0.8433	0.8909	0.8443	0.9170
2	1.0000	0.9335	0.0665	0.9192	0.0629	0.8677	1.0674	0.8685	1.0856
3	1.0000	0.9489	0.0511	0.9466	0.0492	0.8984	1.0686	0.8986	1.0750
4	0.9307	0.9215	0.0092	0.9182	0.0088	0.9131	0.9448	0.9149	0.9632
5	1.0000	0.9377	0.0623	0.9251	0.0588	0.8762	1.0671	0.8768	1.0862
6	0.9071	0.8945	0.0126	0.8875	0.0148	0.8824	0.9350	0.8840	0.9653
7	0.8915	0.8793	0.0122	0.8717	0.0191	0.8676	0.9432	0.8704	0.9857
8	0.8210	0.8111	0.0099	0.8065	0.0116	0.8018	0.8446	0.8029	0.8826
9	0.8892	0.8657	0.0235	0.8490	0.0312	0.8427	0.9501	0.8444	0.9852
10	0.8469	0.8410	0.0059	0.8394	0.0054	0.8357	0.8565	0.8369	0.8720
11	0.9534	0.9462	0.0072	0.9430	0.0098	0.9394	0.9767	0.9406	1.0302
12	1.0000	0.9365	0.0635	0.9221	0.0603	0.8738	1.0547	0.8743	1.0812
13	0.9602	0.9479	0.0123	0.9426	0.0135	0.9363	0.9821	0.9389	1.0289
14	1.0000	0.9281	0.0719	0.9085	0.0692	0.8570	1.0789	0.8576	1.0837
15	1.0000	0.9360	0.0640	0.9157	0.0654	0.8727	1.0811	0.8734	1.0977
16	0.8885	0.8806	0.0079	0.8778	0.0084	0.8732	0.9059	0.8749	0.9245
17	1.0000	0.9406	0.0594	0.9265	0.0577	0.8818	1.0632	0.8823	1.0874
18	1.0000	0.9461	0.0539	0.9392	0.0507	0.8929	1.0584	0.8931	1.0703
19	0.9441	0.9366	0.0075	0.9342	0.0073	0.9297	0.9574	0.9308	0.9737

Table 3: Bootstrap with Bandwidth h= 0.028

						2.5%	97.5%	2.5%	97.5%
k	$\hat{\theta}_k$	$\tilde{\theta}_k$	bias$_k$	Median of $\tilde{\theta}^*_{k,b}$	Std. Dev.	Bias Corrected		Centered on $\tilde{\theta}_k$	
1	0.8692	0.8452	0.0240	0.8429	0.0151	0.8248	0.8792	0.8258	0.8856
2	1.0000	0.9255	0.0745	0.9060	0.0594	0.8554	1.0534	0.8586	1.0711
3	1.0000	0.9400	0.0600	0.9330	0.0448	0.8847	1.0469	0.8858	1.0591
4	0.9307	0.9106	0.0201	0.9084	0.0107	0.8953	0.9369	0.8967	0.9415
5	1.0000	0.9298	0.0702	0.9133	0.0551	0.8640	1.0466	0.8670	1.0734
6	0.9071	0.8846	0.0226	0.8805	0.0153	0.8659	0.9219	0.8683	0.9393
7	0.8915	0.8697	0.0219	0.8639	0.0182	0.8512	0.9232	0.8543	0.9563
8	0.8210	0.8015	0.0195	0.7993	0.0125	0.7855	0.8323	0.7867	0.8410
9	0.8892	0.8563	0.0329	0.8465	0.0289	0.8272	0.9330	0.8309	0.9724
10	0.8469	0.8317	0.0152	0.8303	0.0079	0.8204	0.8500	0.8213	0.8547
11	0.9534	0.9357	0.0177	0.9334	0.0115	0.9215	0.9638	0.9228	0.9789
12	1.0000	0.9277	0.0723	0.9114	0.0577	0.8601	1.0458	0.8623	1.0657
13	0.9602	0.9360	0.0242	0.9329	0.0153	0.9158	0.9701	0.9178	0.9794
14	1.0000	0.9213	0.0787	0.9011	0.0660	0.8470	1.0774	0.8499	1.0994
15	1.0000	0.9279	0.0721	0.9053	0.0627	0.8602	1.0767	0.8629	1.1178
16	0.8885	0.8703	0.0181	0.8680	0.0105	0.8562	0.8965	0.8582	0.9052
17	1.0000	0.9328	0.0672	0.9150	0.0538	0.8704	1.0471	0.8722	1.0738
18	1.0000	0.9358	0.0642	0.9250	0.0488	0.8764	1.0468	0.8778	1.0620
19	0.9441	0.9263	0.0177	0.9244	0.0100	0.9126	0.9504	0.9137	0.9550

The results obtained by bootstrapping process the data differences are negligible**.** The difference in the results of the random selection process, the data is normally distributed.

RESULTS AND DISCUSSION

Data Envelopment Analysis (DEA) is a non parametric method for calculating the size of performance of a group of units which have the same activity. The obtained boundary using this method is an accessible relative boundary in the real world. Unlike parametric method in which units are measured according to a boundary which is generally inaccessible in the real world, the efficiency obtained from this method was a relative amount not its real quantity. In other words, the efficiency obtained from this method was a projection of its real quantity. Due to the uncertainty of the distribution of the precision of the estimated effectiveness is questionable. Introduced by Efron (1979), Bootstrap simulation process can be used to enhance the accuracy of the estimated efficiency. In this paper we provided a deep insight in the workings of Bootstrap DEA and we addressed the important issue of implementing hypothesis testing using Bootstrapped efficiency scores. We introduced a procedure for hypothesis testing which may be applied universally and we explained its associated limitations, while we proposed ways to deal with them. Finally, we used our theoretically consistent procedure to construct confidence intervals which serve as acceptance regions of the null hypothesis of no significant difference in efficiency scores. The paper serves as a guide for the users of bootstrap DEA and as a complement of the Simar and Wilson's (1998) paper, especially when hypothesis tests need to be carried out.

ACKNOWLEDGEMENT

We are grateful to Islamic Azad University, Tabriz branch authorities, for their useful collaboration.

REFERENCES

Bahari Hossaini and Habibi Nia (2011). Using bootstrap process in data envelopment analysis models, *Third National Conference on Data Envelopment Analysis*.

Anders H (1998). *A History of Mathematical Statistics* (John Wiley and Sons) New York.

Banker RD (1993). Maximum likelihood, consistency and data envelopment analysis: *statistical foundation. Management Science* **39** 1265-1273.

Charnes A, Cooper WW and Rhodes E (1978). Measuring the Inefficiency of Decision Making Units. *European Journal of Operational Research* **2** 429-444.
Efron B (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics* **9** 1-26.
Efron B (1982). The jackknife, the bootstrap and other resampling plans, *CBMS* **38** SIAM-NSF.
Efron B (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association* **82**(397) 171-185.
Efron B and Tibshirani RJ (1993). *An Introduction to the Bootstrap* (Chapman and Hall) London.
Fare R and Grosskopf S (1985). A nonparametric cost approaches scale efficiency. *Scandinavian Journal of Economics* **87** 594-604.
Farrell MJ (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society, Series A* **120** 253-281.
Ferrier GD and Hirschberg JG (1997). Bootstrapping confidence intervals for linear programming efficiency scores: With an illustration using Italian bank data. *Journal of Productivity Analysis* **8** 19-33.
Korostelev A, Simar L and Tsybakov AB (1995a). Efficient estimation of monotone boundaries. *The Annals of Statistics* **23** 476-489.
Korostelev A, Simar L and Tsybakov AB (1995b). *On Estimation of Monotone and Convex Boundaries* (Pub. Inst. Stat. Univ. Paris) **XXXIX** 18-31.
Simar L (1992). Estimating efficiencies from frontier models with panel data: A comparison of parametric, non-parametric and semi-parametric methods with bootstrapping. *Journal of Productivity Analysis* **3** 167-203.
Simar L and Wilson WP (1998). Sensitivity analysis of efficiency scores: how to bootstrap in nonparametric frontier models. *Management Science* **44**(1) 49-61.
Simar L and Wilson PW (1999c). Estimating and Bootstrapping Malmquist Indices. *European Journal of Operational Research* **115** 459-471.
Simar L and Wilson PW (2000a). Statistical inference in nonparametric frontier models: the state of the art. *Journal of Productivity Analysis* **13** 49-78.
Simar L and Wilson PW (2000b). A general methodology for bootstrapping in nonparametric frontier models. *Journal of Applied Statistics* **27**(6) 779-802.
Simar L and Wilson PW (2001). Testing restrictions in nonparametric frontier models. *Communications in Statistics: Simulation and Computation* **30** 161-186.
Simar L and Wilson WP (2004). Performance of the bootstrap for DEA estimators and iterating the principle. In: *Handbook on Data Envelopment Analysis*, edited by Cooper WW, Seiford ML and Zhu J (Kluwer Academic Publishers) 265-298.
Simar L and Wilson WP (2007). Estimation and inference in two-stage, semi-parametric models of production processes. *Journal of Econometrics* **136** 31-64.
Simar L and Wilson WP (2008). Statistical inference in non-parametric frontier models. In: *The Measurement of Productive Efficiency and Productivity Growth*, edited by Fried OH, Lovell CAK and Schmidt SS (Oxford University Press) Oxford: New York 421-521.
Sajadi SJ and Omrani H (2009). A bootstrapped robust data envelopment analysis model for efficiency estimating of telecommunication companies in Iran. Telecommunications Policy.