Available online at http://ijim.srbiau.ac.ir/
Int. J. Industrial Mathematics (ISSN 2008-5621)
Vol. 9, No. 4, 2017 Article ID IJIM-00904, 14 pages
Research Article

Impact of Outliers in Data Envelopment Analysis

A. Gholam Abri *†

Received Date: 2016-07-01 Revised Date: 2016-12-19 Accepted Date: 2017-02-13

Abstract

This paper will examine the relationship between "Data Envelopment Analysis" and a statistical concept "Outlier". Data envelopment analysis (DEA) is a method for estimating the relative efficiency of decision making units (DMUs) having similar tasks in a production system by multiple inputs to produce multiple outputs. An important issue in statistics is to identify the outliers. In this paper, we attempt to investigate the concept of the outliers determination by data envelopment analysis and assess the manner of decision making units when a sample contains an outlier. We will start by providing a review literature. We will then proceed with our proposed method and discuss the strengths and weaknesses of our method. We will provide some numerical results to demonstrate the applicability of our method.

Keywords : Data Envelopment Analysis (DEA); Statistics; Outlier; Efficiency; Normal Distribution; Production Possibility Set (PPS).

1 Introduction

In recent years, applying data envelopment analysis has been investigated to different and various sciences and vice versa. In this research, we are going to pay attention to a statistical concept in data envelopment analysis.

As we know, the interdecile or interquartile ranges may be used to represent the variability infrequently. A researcher can select either of these measures to represent the variability. Sometimes, there are a few extreme scores in a special distribution that the researcher likes to omit them. Such extreme scores are referred to as "outliers". Specifically, an outlier is a score within a set of data which is so extreme that, by all appearances, it is not representative of the population from which the sample is ostensibly derived. Since, the presence of outliers can dramatically affect the variability as well as the value of the sample mean, it may cause a researcher to believe that the variability of a distribution might best be expressed through the use of the interdecile or interquartile range. That is, when outliers are present, the sample median is more likely than the mean to be a representative measure of central tendency [18].

On the other hand, traditional models in DEA evaluate the efficiency of a group of DMUs, including itself. Against, the super-efficiency model in DEA excludes each DMU from its own reference set, therefore, it is possible to obtain efficiency scores that exceed one.

Banker and Gifford [3], proposed the use of super-efficiency model to screen out DMUs by gross data errors, and obtain more reliable efficiency estimates after omitting the outliers identified. Banker-Gifford method, BG hereafter, is sketched

*Corresponding author. Amirgholamabri@gmail.com, Tel: +98912 6034180
†Department of Mathematics, Firoozkooh Branch, Islamic Azad University, Firoozkooh, Iran.

for when some DMUs contaminated and, with the result that, classified erroneously as efficient. The first step in BG recognizes the outliers as DMUs whose super-efficiency score exceeds a pre-specified screen level. In the next step, DMUs identified as outliers are removed, and a traditional DEA model such as BCC or CCR model will be used by the remaining DMUs. They prove that the correlations between real efficiency and estimated super-efficiency are negative for the subset of efficient DMUs. Indeed, Banker and Gifford's procedure [3] for identifying outliers generalizes Timmer's procedure [21].

Timmer proposes discarding a specific percentage of efficient DMUs from PPS and re-estimating the production frontier by using the remaining DMUs. Another way to explain Timmer's procedure is that a specific proportion of efficient DMUs are classified as outliers and excluded before re-estimating the efficiency of the remaining DMUs. BG's procedure differs from that of Timmer's in that they propose the use of a screen based on the super-efficiency score to identify DMUs more likely to be contaminated by noise. In other words, instead of throwing out an arbitrary set of efficient DMUs, BG suggest that only DMUs by super-efficiency scores higher than a pre-selected screen should be omitted.

Subsequently, Banker and Change [4] conducted simultaneous experiments for outlier recognition based on the super-efficiency procedure. Note that, the difference between the above super-efficiency model and BCC traditional model is that, when the super-efficiency model is used, the DMU_o under evaluation will not be included in the reference set for the constraints. Since DMU_o under evaluation is eliminated from the reference set in the super-efficiency model, we are not sure that a convex combination can be generated from the remaining DMUs to envelop DMU_o from under for inputs and from above for its outputs. Banker and Gifford [3] proved that, while there is a feasible solution for the super-efficiency model of CCR characteristic, it may not be a feasible solution for the super-efficiency model of BBC characteristic of certain extreme DMUs. Though, for avoiding the computational problem associated with infeasible programs of BCC super-efficiency model, Banker and Gifford [3] suggested a modified model. But, the method developed has two basic difficulties. Firstly, sometimes, when we contaminate some units, only a few of them are identified as outliers (Bellini. [6]). Secondly, it identifies those DMUs as outliers when super-efficiency exceeds a predefined measure. Whereas, probably, in a set of data, some units are extremely efficient and their efficiency scores are higher than pre-specified level by super-efficiency model and this data set has no outlier at all. That is, indeed, the first problem of this model is that some outliers are hidden but not identified as outliers. The second is that some units are not outlier but identified as outlier erroneously.

The previous outlier identification methods were only related to overly efficient outliers. But, only Johnson and McGinnis [15] used the concept of "inefficient frontier" to detect possible outliers performed poorly. Unfortunately, "inefficient frontier" is ad hoc and inconsistent with the DEA standard axioms. Production theory presumes that DMUs are bounded to those by superior performance and that of interior points respected to the efficient frontier and always feasible. Simply applying former procedure, e.g., Pastor et al. [16] to inefficient DMUs violates standard axioms of DEA (production theory) and thus, it is logically problematic.

Chen and Johnson [8] developed a unified model for identifying both inefficient and efficient outliers in DEA. Moreover, by allowing to detect the outliers, the method described is compatible with a relaxed set of DEA axioms. That is, the adaptation of a relaxed set of DEA axioms permits the identification and ranking of both efficient outliers influencing the efficiency evaluations and inefficient outliers probably affecting post analysis procedures. But, one of the most important problem of this method is that, it is relaxing the assumption of free disposability. In other word, this method violates DEA standard axioms. Besides, the applications of the model suggested may be problematic, if the data set is ill-conditioned, i.e., the number of DMUs is small and the variables do not vary over a sufficiently wide range.

At last, Bellini [6] combined the super-efficiency DEA and the forward search to identify the outliers. The forward search is a statistical method originally introduced in linear and nonlinear regression by Atkinson and Riani [1] and so, it is the first effort to extend the approach to a linear programming technique. As we know, there

are several methods for outlier identification dividing the data set into two parts: a clean subset and an outlier subset. The clean data are used to evaluate the model parameters. By contrast, the forward search is based on the following idea: they are built up of an initial subset of a few DMUs adding an additional DMU, step by step. The increasing subset is built up by those DMUs that are the closest according to a pre-specified level. But, this method has some difficulties as follows: when we note to a statistical model, for instance a regression model, we can evaluate its parameters on a subset of data and calculate fitting errors. When focusing on DEA, we can calculate efficiency scores which are not parameters for fitting the model to data. Thus, fitting errors can not be calculated. The next one is the definition of boundaries by which the inference on outlier will be made. Additionally, the calculation process is very complex and the simulation developed in that paper is for when there is one output and multi inputs. Of course, theoretically, the process of applying the method is represented in a case by multi inputs and multi outputs. But, practically, the method developed is very difficult and complex.

In continue, this paper proceeds as follows: Section 2 discusses the basic DEA models, defining the outliers and reviewing some methods to find them. Section 3 scrutinizes the concept of outlier and proposes a new method for finding it by complementary discission in data envelopment analysis. Section 4 provides some numerical examples according to what said in previous sections. Finally, conclusions are given in section 5.

2 Background

Data Envelopment Analysis (DEA) is a technique being used widely in the literature of supply chain management. This non-parametric, multi-factor approach enhances our ability to capture the multi-dimensionality of the performance discussed earlier. More formally, DEA is a mathematical programming technique to measure the relative efficiency of decision making units (DMUs) where each DMU has a set of inputs to produce a set of outputs, (Ross et al. [17]). Consider $DMU_j, (j = 1, ..., n)$, where each DMU consumes m inputs to produce s outputs. Suppose that the observed input and output vectors of DMU_j are $X_j = (x_{1j}, ..., x_{mj})$ and $Y_j = (y_{1j}, ..., y_{sj})$ respectively, and let $X_j \geq 0$, $X_j \neq 0$, $Y_j \geq 0$, and $Y_j \neq 0$.

The production possibility set T_c is defined as:

$$T_c = \Big\{(X,Y) \mid \ X \geq \sum_{j=1}^{n} \lambda_j X_j, \ Y \leq \sum_{j=1}^{n} \lambda_j Y_j \ \Big\}$$

$$where \quad \lambda_j \geq 0, \quad j = 1, \ldots, n$$

The above definition implies that the CCR model is as follows, (Charnes et al. [7]):

$$\begin{aligned} &Min \quad \theta \\ &s.t \\ &\sum_{j=1}^{n} \lambda_j x_{ij} \leq \theta x_{io}, \qquad i = 1, \ldots, m \\ &\sum_{j=1}^{n} \lambda_j y_{rj} \geq y_{ro}, \qquad r = 1, \ldots, s \\ &\lambda_j \geq 0, \qquad\qquad j = 1, \ldots, n \end{aligned} \tag{2.1}$$

Moreover, the production possibility set T_v is defined as:

$$T_v = \Big\{(X,Y) \mid \ X \geq \sum_{j=1}^{n} \lambda_j X_j, \ Y \leq \sum_{j=1}^{n} \lambda_j Y_j \Big\}$$

$$s.t \quad \sum_{j=1}^{n} \lambda_j = 1, \lambda_j \geq 0, \quad j = 1, \ldots, n$$

The above definition implies that BCC model is as following, (Banker et al. [2]):

$$\begin{aligned} &Min \quad \theta \\ &s.t \\ &\sum_{j=1}^{n} \lambda_j x_{ij} \leq \theta x_{io}, \qquad i = 1, \ldots, m \\ &\sum_{j=1}^{n} \lambda_j y_{rj} \geq y_{ro}, \qquad r = 1, \ldots, s \\ &\sum_{j=1}^{n} \lambda_j = 1 \quad \lambda_j \geq 0, \qquad j = 1, \ldots, n \end{aligned} \tag{2.2}$$

The efficiency data envelopment analysis models assessing decision making units are unable to discriminate between efficient DMUs. The discrimination of these efficient units is an interesting subject matter. For ranking decision making units, an important model is proposed by Ander-

sen and Petersen (AP). This model is:

$$\begin{aligned} & AP: Min \quad \theta_o \\ & s.t. \\ & \sum_{j=1,j\neq o}^{n} \lambda_j x_{ij} \leq \theta_o x_{io} \quad ,i=1,\cdots,m \\ & \sum_{j=1,j\neq o}^{n} \lambda_j y_{rj} \geq y_{ro} \quad ,r=1,\cdots,s \\ & \lambda_j \geq 0 \quad ,j=1,\cdots,n, j\neq o \end{aligned} \tag{2.3}$$

Definition 2.1 ***(Reference Set)*** *For a DMU_o, the reference set E_o will be:*
$E_o = \{j \mid \lambda_j^* > 0$, in some optimal solution to model (1) or (2)$\}$ (Cooper et al., [9]).

Definition 2.2 ***(Pareto-Koopmans Efficiency)****. A DMU is fully efficient, if and only if, it is not possible to improve any input or output without worsening some other input or output, (Cooper et al., 2002).*

Definition 2.3 *A DMU_o is extreme efficient, if and only if it satisfies the following two conditions:*
(i) It is efficient(Pareto-Koopmans Efficient).
(ii) $\mid E_o \mid = 1$. (Gholam Abri et al. [12])

Definition 2.4 ***(Median)*** *The median is the middle score in a distribution. In order to determine the median, if there is an odd number of scores in a distribution, the following procedure can be used: Divide the total number of scores to 2 and add 0.5 to the result of the division.*

The value calculated shows the ordinal position of the score representing the median of the distribution. Thus, if we have a distribution consisted of five scores (e.g., 6,8,9,13,16), we will divide the number of scores in the distribution to two, and add 0.5 to the result of the division. The obtained value 3 represents that if five scores are arranged ordinally, the median will be the 3rd score in the distribution. Respected to the mentioned distribution, the value of the median will be equal to 9, since it is the score in the 3rd ordinal position. If there are an even number of scores in a distribution, there will be two middle scores. The median is the average of the two middle scores (Sheskin., [18]).

Theorem 2.1 ***(Central Limit Theorem)*** *If a sample by n member is taken from a statistical population by the mean of μ and the variance of σ^2, if n is adequately large ($n \geq 30$), then the sample average $\overline{X}$ will have a normal distribution by the mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.*

Here, the concept of outliers and the way of recognizing them will be discussed. An outlier is an observation or subset of observations in a set of data appearing inconsistent with the rest of the data. In most instances, inconsistency is reflected in the magnitude of an observation. That is, it is either much higher or much lower than any of the other observations (Banker et al., [4]). Yet, what appears to be an inconsistent/extreme score to one observer may not appear to be so to another. Barnett and Lewis [5] emphasizes that a definable characteristic of an outlier is that it elicits genuine surprise in an observer.
To illustrate the fact that what may surprise an observer may not surprise another, we will consider an example cited by Barnett and Lewis [5]. They represented data developed by Fisher, Corbet, and Williams [10] describing the number of moths of a specific species caught in light-traps mounted in a geographical locale in England. The following 15 observations were obtained.

3,3,4,5,7,11,12,15,18,24,51,54,84,120,560

Barnett and Lewis [5] pointed out that, though the value 560 might appear to be an observation surprising most observers, in fact, it is not an anomaly. The reason why 560 would not be classified as an outlier, is that an experienced entomologist would be privy to the fact that the distribution under study is characterized by a marked skewness and consequently, an occasional extreme score in the upper tail such as the value 560 is a matter-of-fact occurrence. Thus, a researcher familiar with the phenomenon under study would not classify 560 as an outlier. Stevens [20] represented that there are basically two strategies used to deal with outliers. The first strategy is to develop and apply procedures for identifying outliers. By the latter strategy, criteria should be established to determine under what conditions one or more scores identified as outliers should be deleted from a set of data. The second approach is to develop statistical procedures not influenced or least affected by the

presence of outliers. Such procedures commonly called robust statistical procedures. The term "robust" as noted earlier, refers to procedures not overly depended on critical assumptions regarding an underlying population distribution. The discussion of outliers within the framework of robustness is predicated on the fact that the presence of outliers may lead to the violation of one or more assumptions underlying a statistical test.

Barnett and Lewis [5], by representing the most comprehensive source of the subject, describe a large number of tests for identifying outliers. They describe 48 tests alone for the detection of outliers in data assumed to be drawn from a normal distribution. Some tests are designed to identify a single outlier. Others are designed to identify multiple outliers and the rest are specific respected to the identification of outliers in one or both tails of a distribution. Additionally, tests are described to detect outliers in data assumed to be drawn from any number of a variety of non-normal distributions. Given the large number of tests available for detecting outliers, it is not unusual that two or more tests applied to the same set of data may not agree with one another whether a specific DMU should be classified as an outlier or not. In order to recognize the outliers, there are some methods. From which two of them are reviewed in following:

Two Procedures for Identifying Outliers.

First. Sprent [19],[22] proposes a procedure for identifying outliers as being relatively robust. The procedure uses equation (2.4) to determine whether a score in a sample of n DMUs should be classified as an outlier or not:

$$\frac{|X_i - M|}{MAD} > Max \tag{2.4}$$

Where:

X_i represents any of the evaluated scores n respected to whether it is an outlier.

M is the median of the n scores in the sample.

MAD is the median absolute deviation.

Max is the critical value the result to the left of the inequality must exceed to conclude the value X_i is an outlier.

To illustrate the application of equation (2.4), assume we have a sample consisted of the following five scores: 2,3,4,7,18. Next, we determine the median or the middle score of the sample as 4. We compute the absolute deviation of each score from the median: $|2-4|= 2$, $|3-4|= 1$, $|4-4|= 0$, $|7-4|= 3$, $|18-4|= 14$. By arranging four deviation scores ordinally (0,1,2,3,14), we determine the median of five deviation scores as 2. The latter value represents MAD in equation (2.4). Since, the only value we would suspect to be an outlier is the score 18, we use that value to represent X_i in equation (2.4). We will assume the data are derived from a normal distribution, so we apply the value Max=5. Substituting the appropriate values in the left side of equation (2.4), we compute $\frac{|18-4|}{2} = 7$. The obtained value 7 is greater than Max=5, so we conclude that the score 18 is an outlier.

Second. The latter test is developed by Grubbs (1969) as follows:

$$T_n = \frac{|\ X_i - \overline{X}\ |}{\overline{s}} \tag{2.5}$$

Calculated test of statistic T_n is referred to the "*extreme studentized residual*". By using the equation (2.5) to compute T_n, we require a sample mean (based on all scores including the suspected outlier) to be subtracted from the value of a suspected outlier (X_i).

The resulting difference is divided to by the value calculated for $\overline{s}$ which is based on all scores in the sample, just like the mean including the suspected outlier with the equation following:

$$\overline{s} = \sqrt{\frac{\sum(X_i - \overline{X})^2}{n-1}} \tag{3}$$

The value calculated for T_n evaluated by special tables may be found in resources describing the test. In order to illustrate the application of equation (2.5), assume we have a sample consisted of the following five scores:

2,3,4,7,18.

By using equation (2.5), according to data mentioned $T_n = \frac{|18-6.8|}{6.53} = 1.72$ which its turn-out is significant at the level 0.05. So, we conclude that the score 18 is an outlier,

(Sheskin., 2000).

3 Evaluating DMUs in the Presence of the Outliers and Complementary Discussion

3.1 Proposed Method

In this section, a statistical method is used to recognize outliers. When an inferential statistical test is used by one or more samples to draw inferences of one or more populations, such a test may make certain assumptions about the shape of the distribution of underlying population. The most commonly used assumption is that a distribution is normal. That is, the most important continuous probability distribution is normal one in the whole statistics. Its histogram is called "Normal Curve" and it has a bell-shape representing some different and numerous events in the nature, industries and researches.

In the year 1733, Demoivre found out a mathematical formula for normal distribution. Normal distribution parameters include the average μ, ($-\infty < \mu < +\infty$) and the variation $\sigma^2 > 0$. There is a simple interpretation about standard deviation in normal distribution explained by Fig. 1. Any normal distribution can be converted into what is referred to as the standard normal distribution by assigning it (μ=0, σ=1).

Standard normal distribution is used more frequently in inferential statistics than any other theoretical probability distribution. The use of the term theoretical probability distribution in this context is based on the fact that it is known that, in the standard normal distribution, a certain proportion of cases will always fall within specific areas of the curve. As a result, if one knows how far a score is removed from the mean of the distribution, one can specify the proportion of cases obtaining that score as well as the likelihood of randomly selecting or objecting by the score. So, based on the normal distribution explained previously, central limit theorem and that, experimentally, the number of variables by normal distribution is usually more than variables by non-normal distribution, identification of outliers is also relied on this distribution. As explained in previous part, an outlier is the data by values greater or smaller than the others. On the other hand, data envelopment analysis, is a method for evaluating the efficiency of decision making units and identifying the efficient frontiers. Suppose that all decision making units are evaluated in a specific example and an efficiency score would be concluded for each of them. The score is between 0 and 1, (Cooper et al. [9]). In addition, suppose that all identification criteria of outlier units are the score of units efficiency. Because, the units by lower efficiency scores have no effect in assessing the units by upper efficiency, no attention will be paid to the outliers by lowest efficiency scores when evaluating decision making units (weak outlier). This subject is the main difference between the definition of outliers in DEA and in Statistics.

In this way, the outliers are considered in such

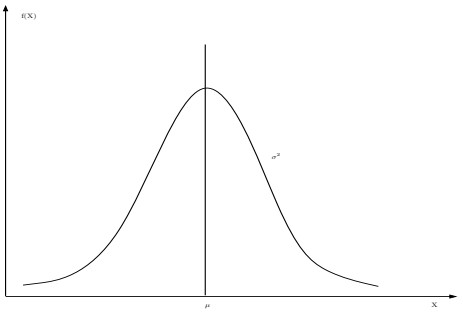

Figure 1: Normal curve

a way that their efficiency scores are absolutely 1 (strong outlier). It is expected that, in samples by such outliers, the efficiency score to be 1 and the rest of the units, in comparison with the score 1, represents a sudden decrease. In these samples, a few outliers units have the efficiency score 1, the rest would be inefficient by a great difference. So, it seems that the evaluation of the units would be unreal in such circumstances. It is why some outliers are exception in population considered and it is not logical to compare common units with them.

For instances, Albert Einstein and Isaac Newton were very famous and popular for the best contribution in human society. But, as we know, these two scientists, in comparison with other people, are the exceptions and if we compare them, all people will be inefficient. So, these two people are as outliers. Therefore, prior to proposing our method for identifying outliers, a definition of them in DEA is provided.

Definition 3.1 ***(Outlier in DEA)*** *DMU is an outlier, if and only if, it's efficiency score is abso-*

lutely 1 and the efficiency score of other units in comparison to them have a sudden decrease and close to 0. In other words, outliers have 2 characteristics:

I. They are recognized as super-efficient units.
II. The efficiency score of other units compared to them are extremely low.

Initially, the outliers would be determined and omitted from the sample considered, then the rest of units would be evaluated to be closer to the reality. So, the score concluded for the rest of the units will be considered as efficiency scores.

As we know, the main objective is to determine the outliers by data envelopment analysis and the methods mentioned suggest many outliers appearing as an outlier according to the observation by data envelopment analysis are not really so. To overcome this difficulty, a new method is proposed for recognizing real outliers as follows.

1. First, a sample by one main model in DEA such as CCR, BCC or etc is used and its efficiency scores will be calculated.

2. Second, Kolmogrov-Smirnov test is done on the units efficiency scores. Also, the histogram of the units efficiency scores can confirm the data having or not having normal distribution. If the test confirms that the units efficiency scores have a normal distribution, then, it can be said that this sample has no outlier. But, if it figures out no normal distribution, the sample will have outlier. In this step, all units having the efficiency score 1, are wiped out from the PPS and then, the rest of units will be evaluated again by a DEA model. Again, Kolmogrov-Smirnov test is used to consider the properties of the efficiency scores of the units. If the test demonstrates that the units efficiency scores have a normal distribution, it can be claimed that the new sample has no outlier anymore and the units efficiency scores will be accepted without any difficulties. Nevertheless, the procedure goes on to determine and omit all samples outliers.

3. After omitting all outliers, the efficiency scores obtained would be considered as units efficiency scores in the next step. By doing so, the real efficiency of the units will be observed.

Here, an important point is the determination of efficiency score of outlier units removed from the PPS. Because, based on the reasons explained, these units are removed to estimate the real efficiency score of other units. But, there is no method for calculating the efficiency score of these units.

4. Now, to calculate the efficiency scores of the outliers compared to other DMUs the following method will be used. Firstly, according to what said previously, outliers will be identified if existing. Then, in order to help our development, we classify the set of n DMUs into two classes:

1. Class Ω_1, a clean subset containing all common units without any outlier.
2. Class Ω_2, an outlier subset containing all outliers. (if existing).

If the second set is vacuous, that is, when production possibility set is without outlier, the evaluation of the units will not be difficult. But, when the second set is not vacuous, having at least one member, for evaluating the first set, after removing the outlier units, the method will be used as said before. The evaluation of the second set can be as following. Assume the set Ω_2 contains outliers as:

$\Omega_2 = \{DMU_1, \ldots, DMU_l\}$

So, we calculate the efficiency scores of the second set member as follows:

First, add each member of Ω_2 to Ω_1 one by one:

$\Gamma_1 = \Omega_1 \bigcup \{DMU_1\}$

$\Gamma_2 = \Omega_1 \bigcup \{DMU_2\}$

$\vdots$

$\Gamma_l = \Omega_1 \bigcup \{DMU_l\}$

Then, we use the AP model for Γ_i, for each i $\in \{1, 2, \ldots, l\}$ separately, and we will obtain the efficiency scores for DMU_1, DMU_2, ..., DMU_l introduced as outliers (Gholam Abri et al., [11] [13] [14]). An important point is that the efficiency scores of these units surely are far away 1. So, noting that the data are usually of the first set, that is, Ω_1, the efficiency scores will be between 0 and 1. In addition, outliers have the efficiency scores considerably higher than 1 compared to other common units, so the evaluation of the whole units will be closer to the reality. The mentioned method can be represented with the following flowchart: However, in continue, the paper notes to different cases of DMUs evaluation by the method developed. As it will be seen later, it is very important to note to different probable conditions. By paying attention to this subject the probable difficulties of the method proposed will be removed. Of course, it must be said that the evaluation is not very simple in all cases. In other word, there may be some different and ex-

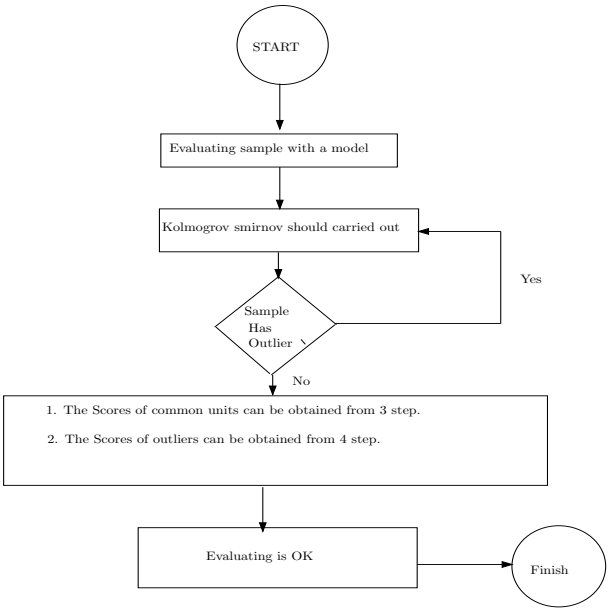

Figure 2: Flowchart of the Method

ceptional conditions. So, it is necessary to have a complementary discussion.

3.2 Complementary Discussion

As mentioned, for the classification of efficiency score of the units under evaluation, the outliers are units that their efficiency score, in comparison with other units, are very small or very large in data envelopment analysis. So, one of the following items may be encountered in the investigation of the outliers existence.

Case 3.2.1

The distribution of efficiency scores is normal without any outlier. An example is considered in Fig. 3. In this case, the units under evaluation

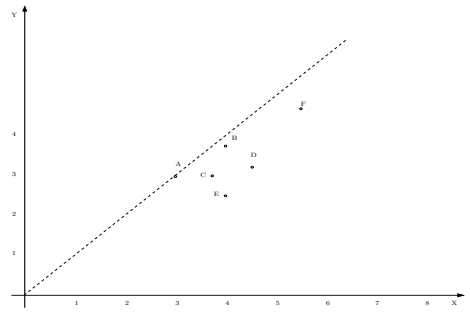

Figure 3: Without Outlier

encounters no problem because of not having any outlier.

Case 3.2.2

The set under-study has only strong outliers as discussed in details.

Case 3.2.3

The set under-study has just weak outliers. Fig. 4 makes the matter clear. It is obvious that

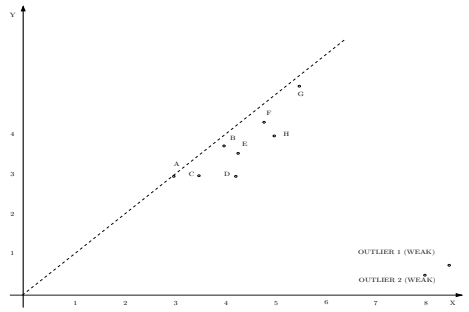

Figure 4: Weak Outlier

Kolmogrov-Smirnov test does not confirm the normal distribution. But, the histogram of the sample represents the data having weak outliers. In this way, since the efficient frontier is characterized by the units having efficiency score 1, these weak outliers have no effect in the evaluation of the rest units. So, the units assessment is done without any omitting.

Case 3.2.4

In this case, our assumption is that there are both strong and weak outliers. The units are distributed in a way that Kolmogrov-Smirnov test verifies normal distribution wrongly.

It represents a sense when the outliers exist in sample, but, because of a symmetric distribution between strong and weak outliers, Kolmogrov-Smirnov test would verify it as a normal distribution. Also, the histogram of the units efficiency scores confirms the matter. In such case, the steps followed are in progress.

At first, the units are evaluated by a basic model in DEA, such as CCR or BCC and then, efficient units are wiped out from the production possibility set. Two cases can be considered:

A. Kolmogrov-Smirnov test and units efficiency

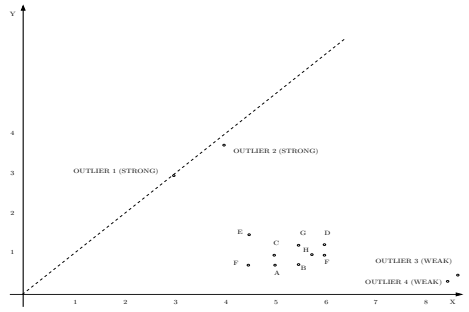

Figure 5: Strong and Weak Outlier

scores histogram also would not figure out the normal distribution any more. In this circumstance, the data distribution are not normal. Because the strong outliers are omitted from the PPS, but there are weak outliers in the sample under-study. In following, after omitting strong outliers, units efficiency scores are calculated again and then it goes to the end.

B. Kolmogrov-Smirnov test and histogram represent a normal distribution. But, both of them claim that the normal distribution has a worse condition compared to previous case. That is, the normal distribution, at least, moves a little away from symmetric condition. In this context, the rest units are evaluated by the help of a main model in DEA such as CCR or BCC and so, the efficient units are omitted from the PPS. In continue, either "A" or "B" will be encountered. If "A", the process will be cleared as we said perviously. If "B", the process will be continued to meet "A". Since, the number of outliers in comparison with the number of other units is few; mostly, 5 percent, omitting the efficient units will be continued till 5 percent of total DMUs.

Important Points.

As said in Fig. 5, because there are both strong and weak outliers in sample and Kolmogrov-Smirnov test verifies it as a normal distribution, after omitting the strong outliers from the PPS and using Kolmogrov-Smirnov test again on the rest units, we will encounter a condition in which the test does not confirm it as a normal distribution for the rest units. That is, by continuing the process, the case "A" would be met certainly. So, after wiping out the strong outliers in this situation, the rest units will meet a circumstance like the case 3.2.3

As it can be seen, the remaining units by weak outliers represents that the data have no normal distribution. In fact, the units remained will cause the conditions like the case 3.2.3 that, in continue, a similar procedure will be applied to.

4 Applications to some data sets

4.1 Data Set 1

Consider 10 DMUs by a single input and output as defined in Table 1. By evaluating these DMUs according to CCR model, we find out:

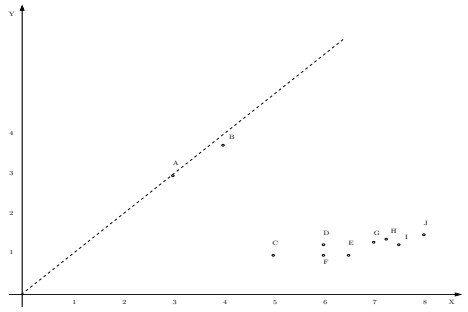

Figure 6: Data set in CCR model

As observed in Fig. 6, DMU_A is efficient and the other units are inefficient. At first, according to the new method, Kolmogrov-Smirnov test is done on 10 units.

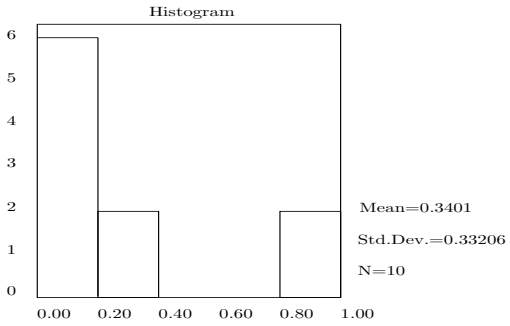

Figure 7: The Histogram of 10 units

The results represent that the sample has an outlier. Moreover, Fig. 7, the histogram of the sample verifies the subject. Then, the unit A which is efficient would be wiped out from the PPS and after that, 9 remained units would be evaluated in accordance with CCR model. It is observed that DMU_B is efficient and the rest units are inefficient.

Kolmogrov-Smirnov test is done again on 9 units remained. Results represent that the rest sample has outlier as well. Again, Fig. 8, the histogram of 9 remained units efficiency scores confirms no normal distribution as follows:

The unit B is wiped out from the new PPS and the 8 remained units are evaluated by CCR model.

In 8 units remained, Kolmogrov-Smirnov test confirms a normal distribution. On the other hand, the histogram represents the matter in

Table 1: Data and Result

DMUs	Input	Output	The efficiency score (All units)	$step_1$ (after omitting A)	$step_2$ (after omitting A,B)
A	3	3	1.0000	——	——
B	4	3.75	0.9375	1.0000	——
C	5	1	0.2000	0.2133	0.9600
D	6	1.25	0.2083	0.2222	1.0000
E	6.5	1	0.1583	0.1641	0.7385
F	6	1	0.1667	0.1778	0.8000
G	7	1.3	0.1857	0.1981	0.8914
H	7.25	1.4	0.1931	0.2060	0.9269
I	7.5	1.25	0.1667	0.1778	0.8000
J	8	1.5	0.1875	0.2000	0.9000

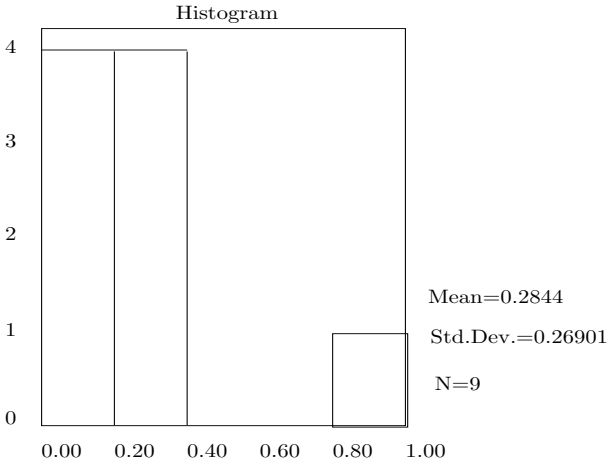

Figure 8: The Histogram of 9 units

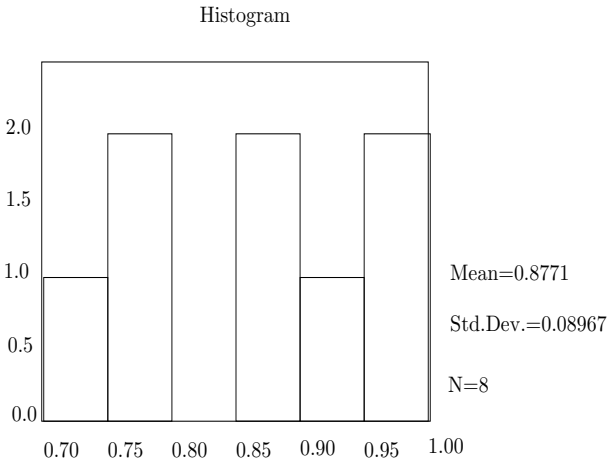

Figure 9: The Histogram of 8 units

Fig. 9.

It is claimed that the units A and B are outliers recognized along with 2 steps. So,we set:
$\Omega_1 = \{C, D, E, F, G, H, I, J\}$, $\Omega_2 = \{A, B\}$
$\Gamma_1 = \Omega_1 \bigcup \{A\}$, $\Gamma_2 = \Omega_1 \bigcup \{B\}$
Then, after omitting outliers, the actual assessment is done and real conclusions are figured out in the last column of Table 1. In continue, the AP model will be used to calculate the efficiency score of A and B for Γ_i, for each i $\in \{A, B\}$. So, the efficiency score of the unit A is $\theta_A^* = 2.67$ and the efficiency score of the unit B is $\theta_B^* = 2.31$. By a little more attention to the results obtained, the difference between the new method and the pervious methods will be determined.

Point. As illustrated in figure 6, if the BCC model was applied to the data set, units A and B, would be recognized as outliers, since the efficient frontier would not have a significant change.

4.2 Data Set 2

. As an application, the approach is used to some branches of Iranian banks. The district is consisted of 50 branches. Each branch uses 4 inputs and like all systems have a process, 4 outputs would be concluded. All data are normalized.
At first, the units will be evaluated by the CCR model. Data and results are summarized in Table 2. As it is represented in the conclusion column, 20th and 30th units are efficient and the rest are not. In addition, the highest efficiency score, except these two units, is the unit 12 that $\theta_{12}^* = 0.4159$. It represents a considerable gap between 0.4159 and 1.
According to the new method, Kolmogrov-Smirnov test is used to these 50 units. It is shown that the data have no normal distribution. The histogram of the units efficiency scores confirms that the data have no normal distribution as shown in Fig. 10:

Now, DMU_{20} and DMU_{30} would be omitted

Table 2: Data and Result

DMUs	Input1	Input2	Input3	Input4	Output1	Output2	Output3	Output4	θ^*_{Old}	θ^*_{New}
1	0.9333	0.7006	0.8276	0.4404	0.0556	0.1077	0.0026	0.5803	0.0586	0.5257
2	0.8667	0.9641	0.6810	0.3793	0.0518	0.0684	0.0107	0.6871	0.0805	0.6832
3	1.0000	0.8862	0.7974	0.3301	0.0857	0.1394	0.0039	0.7942	0.1069	0.6796
4	1.0000	0.4371	1.0000	0.2058	0.2213	0.1611	0.0064	1.0000	0.2160	0.9116
5	0.9333	0.7964	0.7888	0.3601	0.0949	0.0817	0.0102	0.7922	0.0978	0.7247
6	0.9333	1.0000	0.7759	0.4424	0.0644	0.0330	0.0038	0.5137	0.0516	0.4695
7	0.7333	0.5329	0.6250	0.2579	0.0454	0.0890	0.0023	0.5449	0.0939	0.6310
8	0.6667	0.7365	0.5733	0.1102	0.0253	0.0514	0.0050	0.3229	0.1302	0.4133
9	0.6667	0.8563	0.5086	0.2552	0.0698	0.1010	0.0007	0.6304	0.1098	0.8119
10	0.5333	0.3413	0.5086	0.1698	0.0778	0.1401	0.0172	0.4081	0.1068	0.6665
11	0.8000	0.6467	0.6940	0.2675	0.0687	0.0544	0.0011	0.8005	0.1330	0.8474
12	0.6000	0.6228	0.5043	0.0514	0.0710	0.0743	0.0031	0.4810	0.4159	1.0000
13	0.6000	0.6168	0.5129	0.0617	0.0724	0.0447	0.0010	0.3620	0.2608	0.6380
14	0.7333	0.6347	0.6207	0.1260	0.0591	0.0723	0.0011	0.4404	0.1553	0.5100
15	0.6667	0.7725	0.5431	0.0674	0.0537	0.0663	0.0124	0.5679	0.3745	0.9136
16	0.6000	0.4850	0.5302	0.1461	0.0395	0.0532	0.0068	0.4320	0.1314	0.6125
17	0.8000	0.4731	0.7672	0.1828	0.4035	0.1294	0.0026	0.8467	0.2059	0.8879
18	0.4000	0.2575	0.3707	0.0823	0.0323	0.0174	0.0007	0.4249	0.2295	0.8927
19	0.9333	0.6287	0.8405	0.3395	0.1402	0.3263	0.0207	0.6414	0.0840	0.6339
20	0.0600	0.0200	0.0700	0.0400	0.8000	0.7700	0.2500	0.9000	1.0000	——
21	0.7333	0.5868	0.6552	0.1631	0.1791	0.3236	0.0105	0.6076	0.1656	0.7725
22	0.4000	0.3114	0.3319	0.1007	0.0317	0.0299	0.0028	0.3343	0.1475	0.7133
23	0.1100	0.0970	0.0931	0.1070	0.6677	0.5862	0.1553	0.9231	0.3834	1.0000
24	0.4000	0.3713	0.3233	0.1134	0.0296	0.0475	0.0023	0.2335	0.0915	0.5000
25	0.4667	0.2275	0.4828	0.1774	0.0386	0.0256	0.0006	0.2778	0.0696	0.5265
26	0.8000	0.4910	0.7457	0.2512	0.0663	0.0375	0.0011	0.8220	0.1454	0.8629
27	0.8000	0.6707	0.6983	0.1646	0.0803	0.1486	0.0016	0.6497	0.1754	0.6873
28	0.4667	0.3713	0.4138	0.1911	0.0359	0.0430	0.0032	0.2765	0.0643	0.5027
29	0.7333	0.4251	0.6897	0.1183	0.6867	0.2694	0.0028	0.8741	0.3284	1.0000
30	0.0500	0.0300	0.0800	0.0500	0.7500	0.8000	0.2000	0.9150	1.0000	——
31	0.4667	0.3234	0.0995	0.0545	0.0430	0.0362	0.0006	0.5040	0.4110	1.0000
32	0.5333	0.6287	0.4138	0.2839	0.0228	0.1653	0.0015	0.2482	0.0467	0.4759
33	0.8667	0.7066	0.7543	0.2725	0.2847	0.3709	0.0098	0.7874	0.1284	0.8237
34	0.2000	0.1138	0.1810	0.0309	0.0141	0.0125	0.0005	0.1125	0.1618	0.4807
35	0.5333	0.3054	0.5000	0.1354	0.1500	0.2461	0.0154	0.3450	0.1132	0.7085
36	0.3333	0.3174	0.2802	0.1778	0.0319	0.0445	0.0035	0.2240	0.0622	0.5749
37	0.4000	0.3174	0.3405	0.1592	0.0770	0.0352	0.0006	0.3783	0.1056	0.8027
38	0.6000	0.2934	0.5862	0.2140	0.0852	0.0683	0.0009	0.5308	0.1102	0.7818
39	0.5333	0.3713	0.5000	0.1903	0.0805	0.0628	0.0004	0.4393	0.1026	0.6913
40	0.7333	0.3713	0.7241	0.2646	0.2638	0.2923	0.0226	0.6747	0.1133	0.8198
41	0.4667	0.4431	0.3750	0.5637	0.0580	0.0984	0.0018	0.3487	0.0723	0.6389
42	0.5333	0.4431	0.4440	0.0617	0.0327	0.0251	0.0006	0.3712	0.2674	0.6504
43	0.4667	0.3353	0.3966	0.1364	0.0554	0.0909	0.0022	0.5263	0.1715	0.9579
44	0.4667	0.3413	0.4009	0.2264	0.0441	0.0739	0.0074	0.5425	0.1065	0.9910
45	0.6000	0.4192	0.5172	0.1274	0.0562	0.0508	0.0029	0.5473	0.1909	0.7740
46	0.5333	0.2994	0.4741	0.3540	0.0816	0.0490	0.0022	0.5587	0.0917	0.8901
47	0.4667	0.5449	0.4095	0.0886	0.0385	0.0633	0.0012	0.2773	0.1391	0.5029
48	0.4667	0.4910	0.3750	0.4702	0.0384	0.0282	0.0024	0.4197	0.0870	0.7692
49	0.6667	0.3892	0.6336	0.1749	0.0911	0.2292	0.0015	0.7404	0.1881	0.9317
50	0.2667	0.1796	0.2586	0.0720	0.0387	0.0282	0.0005	0.2483	0.1533	0.7810

from PPS and Kolmogrov-Smirnov test will be implemented on the units remained.

By doing so, the rest of 48 samples have a normal distribution. It can be claimed that there is no

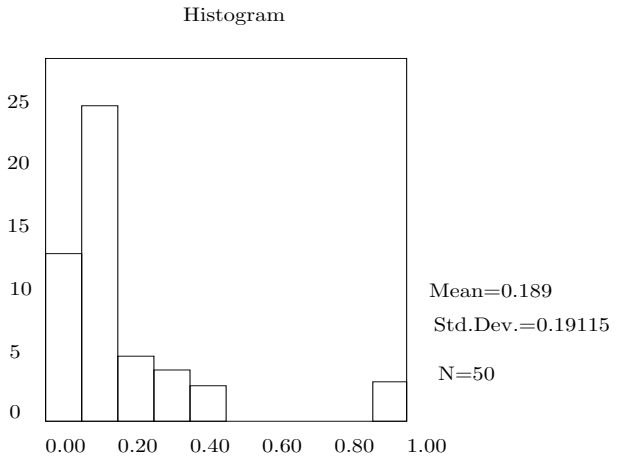

Figure 10: The Histogram of 50 units

outlier in 48 units. On the other hand, Fig. 11, the histogram confirms the matter as following:

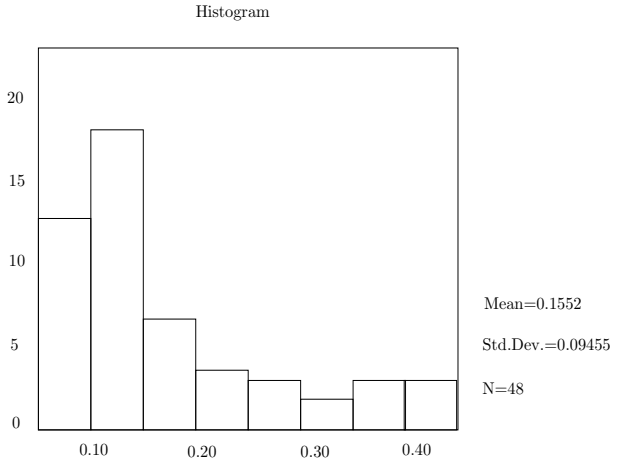

Figure 11: The Histogram of 48 units

Obviously, the units 20 and 30 are efficient as well and have the efficiency score 1. So, these units are outliers. So,we set:
$\Omega_1 = \{DMU_1, ..., DMU_{19}, DMU_{21}$

$$, ..., DMU_{29}, DMU_{31}, ..., DMU_{50}\}$$

$\Omega_2 = \{DMU_{20}, DMU_{30}\}$
$\Gamma_1 = \Omega_1 \bigcup \{DMU_{20}\}$, $\Gamma_2 = \Omega_1 \bigcup \{DMU_{30}\}$
Then, after omitting outliers, the actual assessment is done and real conclusions are figured out in the last column of Table 2. In continue, the AP model will be used to calculate the efficiency score of DMU_{20} and DMU_{30} for Γ_i, for each i $\in \{20, 30\}$. So, the efficiency score of the unit 20 is $\theta^*_{20} = 2.43$ and the efficiency score of the unit 30 is $\theta^*_{30} = 2.38$.
By using the method proposed, real efficiency of the units will be observed. Specially, we can see that the number of efficient units are increased.

5 Conclusion

In this paper, a conceptual application of statistics called "outlier" are studied by data envelopment analysis. Moreover, a new method will be proposed to determine the outliers in samples under evaluation. As a contribution, the units by lower efficiency score have no effect on units by higher scores when evaluating the decision making units and so, the efficiency frontiers would be constructed by efficient units. Then, we are not going to consider the outliers by very small efficiency score.
On the other hand, recognizing the outliers by efficiency score 1 is essential. Because these units are exceptions of the population, the comparison between common units and these outliers leads to the conclusion that there is a long way to the reality. It is anticipated that, after recognizing the outliers and wiping them out from the production possibility set, the conclusion will be closer to the reality.
But, the priority of this method to the previous ones is that, in addition to its very simple computational process like other methods developed formally, it has no need to determine pre-specified level to identify outliers. That is, it works dynamically. Consequently, the problems of the previous methods are solved. According to the new method introduced, there is a problem in practical examples. The outliers by very weak performance (lower efficiency score) in statistical analysis frustrates the outliers effect by a very good performance (higher efficiency score). It makes a symmetric score distributions. In this way, it is possible that Kolmogrov-Smirnov test confirms the distribution scores as normal. But, the outliers with strong performance exist in samples really.

References

[1] A. C. Atkinson, M.Riani, Robust Diagnostic regression analysis, *Springer-verlage*, 2000, New York.

[2] R. D. Banker, A. Charnes, W. W. Cooper, Some models for estimating technical and

scale inefficiencies in data envelopment analysis, *Management Science* 30 (1984) 1078-1092.

[3] R. D. Banker, J. L. Gifford, A relative efficiency model for the evaluation of public health nurse productivity, *Mellon University Mimeo*, 1988, Carnegie.

[4] R. D. Banker, Hsihui Chang, The super-efficiency procedure for outlier identification, not for ranking efficint units, *European Journal of Operational Research* 175 (2006) 1311-1320.

[5] V. Barnett, T. Lewis, Outliers in Statistical Data, 1994 3rd edition. J. Wiley and Sons.

[6] T. Bellini, Forward search outlier detection in data envelopment analysis, *European Journal of Operational Research* 216 (2012) 200-207.

[7] A. Charnes, W. W. Cooper, E. Rhodes, Measuring the efficiency of decision making units, *European Journal of Operation Research* 2 (1978) 429-444.

[8] Wen-Chih Chen, A. L. Johnson, A unified model for detecting efficient and inefficient outliers in data envelopment analysis, *Computers and Operations Research* 37 (2010) 417-425.

[9] W. W. Cooper, L. Seiford, K. Tone, Data envelopment analysis a comprehensive text with Models applications references, DEA solved software, *Third Printing, By Kluwer academic publishers* 2002.

[10] Roland D. Fisher, Graphic Interpretation of the Structures Influencing Mandibular Denture Mucoperipheral Outline Form, The Journal of the American Dental Association 30 (1943) 408-415.

[11] A. Gholam Abri, N. Shoja, M. Fallah Jelodar, Sensitivity and Stability Radius in Data Envelopment Analysis, *Int. J. Industrial mathematics* 1 (2009) 227-234.

[12] A. Gholam Abri, G. R. Jahanshahloo, F. Hosseinzadeh Lotfi, N. Shoja, M. Fallah Jelodar, A New Method for Ranking Non-Extreme Efficient Units in Data Envelopment Analysis, *Optimization Letters* 7 (2013) 309-324.

[13] A. Gholam Abri, An investigation on the Sensitivity and Stability Radius of Returns to Scale and Efficiency in Data Envelopment Analysis, *Applied mathematical modelling* 37 (2013) 1872-1883.

[14] G. R. Jahanshahloo, F. Hosseinzadeh Lotfi, N. Shoja, A. Gholam Abri, M. Fallah Jelodar, K. Jamali Firouzabadi, Sensitivity analysis of inefficient units in Data Envelopment Analysis, *Mathematical and Computer Modelling* 53 (2011) 587-596.

[15] A. L. Johnson, Leon F. McGinnis, Outlier detection in two-stage semi parametric DEA models, *European Journal of Operational Research* 187 (2008) 629-635.

[16] J. T. Pastor, J. L. Ruiz, I. Sirvent, A statistical test for detecting influential observations in DEA, *European Journal of Operational Research* 115 (1999) 542-554.

[17] Anthony Ross, Kathryn Ernstberger, Benchmarking the IT Productivity Paradox: Recent evidence from the manufacturing sector, *Mathematical and Computer Modelling* 44 (2006) 30-42.

[18] David J. Sheskin, Hand book of parametric and nonparametric statistical procedures, *By Chapman and Hall/CRC*, (2000).

[19] Jonathan Sprent, Lifespans of naive, memory and effector lymphocytes, *Current Opinion in Immunology* 5 (1993) 433-438.

[20] J. Stevens, Applied multivariate statistics for the social sciences (3rd ed.). Hillsdale, *NJ: Lawrence Erlbaum Associates*, Publishers 1996.

[21] C. P. Timmer, Using a probabilistic frontier production function to measure technical efficiency, *Journal of Political Economy* 79 (1971) 776-794.

[22] David F. Tough, Jonathan Sprent, Anti-viral immunity: Spotting virus-specific T cells, *Current Biology* 14 (1998) 498 - 501.

Amir Gholam Abri is Associate Professor in Department of Mathematics, Firoozkooh Branch, Islamic Azad University, Firoozkooh, Iran. His research interests are in the areas of Applied Mathematics, Data Envelopment Analysis, Ranking and Sensitivity Analysis. He has published research articles in international journals consist of journals of Elsevier Springer in Data Envelopment Analysis. He is referee of Mathematical journals.