DEA with Missing Data: An Interval Data Assignment Approach

Reza Kazemi Matin[a,*], Roza Azizi[b]

[a] *Associate Professor, Department of Mathematics, Islamic Azad University, Karaj Branch, Karaj, Iran*
[b] *MSc, Department of Mathematics, Islamic Azad University, Karaj Branch, Karaj , Iran*
Received 17 October, 2013; Revised 20 December, 2013; Accepted 02 February, 2014

Abstract

In the classical data envelopment analysis (DEA) models, inputs and outputs are assumed as known variables, and these models cannot deal with unknown amounts of variables directly. In recent years, there are few researches on handling missing data. This paper suggests a new interval based approach to apply missing data, which is the modified version of Kousmanen (2009) approach. First, the proposed approach suggests using an acceptable range for missing inputs and outputs, which is determined by the decision maker (DM). Then, applying the least favourable bounds of missing data along with using the proposed range is suggested in estimating the production frontier. A data set is used to illustrate the approach.
Keywords: Data envelopment analysis, Missing inputs, Missing outputs, Range.

1. Introduction

Data envelopment analysis (DEA) is an approach, which is used for measuring the relative efficiency of a set of decision-making units (DMUs), which convert the same inputs to the same outputs (Charnes et al., 1978). DEA provides efficiency scores and efficient projections for inefficient DMUs. One of the most important qualifications of DEA is that inputs and outputs are known exactly. However, in the real-world DEA applications, there are many cases, in which there are no complete input/output quantities, and data contain some missing values. As a result, the classical DEA models are weak at efficiency evaluation of these kinds of systems.
In recent years, there have been a few studies in the DEA literature dealing with missing data. Kao and Liu (2000) proposed an extended DEA model based on fuzzy theory to handle missing data. They suggested replacing the missing values with a fuzzy number and using the observed data to estimate membership functions of fuzzy efficiency scores. Smirlis et al., (2006) suggested an approach to estimating the amount of missing values. Kuosmanen (2009) in a systematic fashion proposed a method to handle missing data, in which the missed outputs are replaced with zero and missed inputs with a sufficiently large number. This approach will be discussed in the next section in more details, and our new interval data version of this work will be introduced. Azizi (2013) showed some drawbacks of Smirlis (2006) and replaced it by a new approach. Zha et al. (2013) introduced modified DEA models to calculate proper amounts of missing values.

In this paper, after reviewing some approaches to missing data, a connection among some technologies, which can handle missing data, is presented and finally, by using decision makers (DMs) knowledge about the production system, an acceptable range for missing data is assumed, and missing inputs and outputs will be replaced by the worst case in the interval, respectively. The suggested approach can eliminate some of the disadvantages of other approaches.
The rest of the paper is organized as follows. Section 2 is devoted to introducing the required models and notations. In section 3, three approaches in DEA in dealing with missing data will be discussed. The first technique eliminates DMUs with missing values. The second is based on eliminating outputs and inputs with missing values, and the last one is based on taking zero for missing outputs and taking high value for missing inputs. Afterward, we will present our approach based on having a range for missing data in section 4. Some features of this new approach are explored in this section. Section 5 is devoted to presenting some numerical examples. Conclusions are given in section 6.

2. Models and Assumptions

There are some models and notations, which are required in our discussion about missing data.

* Corresponding author Email address: rkmatin@kiau.ac.ir

2.1 Notations

In general, the production possibility set (PPS) in a specific technology could be stated as

$$T=\left\{(\boldsymbol{x},\boldsymbol{y})\in\mathbb{R}_{+}^{m+s}\,\middle|\,\boldsymbol{x}\ \text{can produce}\ \boldsymbol{y}\right\}.$$

To present the required background, assume that there are n DMUs to be evaluated. Each DMU consumes varying amounts of m different inputs to produce s different outputs. Specifically, DMUj consumes amount xij of input i and produces amount yrj of output r. In the classical DEA models, non-negative data is assumed and further it is assumed that each DMU has at least one positive input and one positive output value. Under constant returns to scale (CRS) technology and complete data, the underlying production set could be represented based on data set as

$$T_{DEA}=\left\{(\boldsymbol{x},\boldsymbol{y})\in\mathbb{R}_{+}^{m+s}\,\middle|\,\boldsymbol{x}\geq\sum_j\lambda_j\boldsymbol{x}_j,\boldsymbol{y}\leq\sum_j\lambda_j\boldsymbol{y}_j,\forall j;\lambda_j\geq 0\right\} \quad (1)$$

Furthermore, TDMUo be the production possibility set, obtained by eliminating DMUO from the observe set (DMUO is a unit with missing input and/or output), and TXY be the production possibility set, obtained by eliminating output r' and input i' for all units. These technologies can be introduced as follows:

$$T_{DMUo}=\left\{\begin{array}{l}(x,y)\,\middle|\,x_{io}\geq\sum_{j=1}^{n}\lambda_j x_{ij},y_{ro}\leq\sum_{j=1}^{n}\lambda_j y_{rj};\\ \lambda_j\geq 0,(j\neq o)\\ \lambda_o=0\end{array}\right\} \quad (2)$$

$$T_{XY}=\left\{\begin{array}{l}(x,y)\,\middle|\,x_{i'j}=0,y_{r'j}=0;y_{rj}=\hat{y}_{rj}\ (\forall r\neq r')\\ x_{ij}=\hat{x}_{ij}\ (\forall i\neq i'),(\hat{x}_{ij},\hat{y}_{rj})\in T_{DEA}\end{array}\right\}$$

2.2 CCR model

The following table represents envelopment and multiplier forms in both input and output orients of the CCR DEA model introduced by Charnes, Cooper, and Rhodes (1978) when the unit "o" is under evaluation to get its Farrell (1957) radial efficiency score.

In all the original DEA models, a fundamental assumption is that the inputs and outputs are measured exactly with non-negative values on a ratio scale, and all the data are available. Suppose the value of a specific input/output for at least one unit is unknown, and it is impossible to collect and register complete data set. As a result, using those developed approaches is necessary to deal with missing values in the literature. The next section is devoted to presenting the main approaches in this area and discusses their potential pitfalls in dealing with missing data.

3. Measuring the efficiency score of DMUs with missing data

Several approaches are proposed for dealing with missing data. Here, a classified summary of these approaches is presented.

3.1 Eliminating DMUs with missing values

One of the common strategies in DEA to handle missing data is eliminating DMUs including missing data; see, for example, Neal, Ozcan and Yanqiang (2002) among others. Although this approach is used in some DEA

Input oriented

Envelopment model

$$\begin{aligned}&\text{Min}\quad \theta\\ &\text{s.t.}\quad \textstyle\sum_j x_{ij}\lambda_j\leq\theta x_{io},\ i=1,\ldots,m, \qquad (f.1)\\ &\qquad \textstyle\sum_j y_{rj}\lambda_j\geq y_{ro},\ r=1,\ldots,s,\\ &\qquad \lambda_j\geq 0,\ j=1,\ldots,n.\end{aligned}$$

Multiplier model

$$\begin{aligned}&\text{Max}\quad \theta=\textstyle\sum_j u_r y_{ro}\\ &\text{s.t.}\quad \textstyle\sum_j v_i x_{io}=1, \qquad (f.2)\\ &\qquad \textstyle\sum_j u_r y_{rj}-\sum_j v_i x_{ij}\leq 0,\ j=1,\ldots,n,\\ &\qquad u_r\geq 0,\ v_i\geq 0,\qquad \forall r,\forall i.\end{aligned}$$

output oriented

Envelopment model

$$\begin{aligned}&\text{Max}\quad \varphi\\ &\text{s.t.}\quad \textstyle\sum_j x_{ij}\lambda_j\leq x_{io},\ i=1,\ldots,m, \qquad (f.3)\\ &\qquad \textstyle\sum_j y_{rj}\lambda_j\geq\varphi y_{ro},\ r=1,\ldots,s,\\ &\qquad \lambda_j\geq 0,\ j=1,\ldots,n.\end{aligned}$$

Multiplier model

$$\begin{aligned}&\text{Min}\quad \varphi=\textstyle\sum_j v_i x_{io}\\ &\text{s.t.}\quad \textstyle\sum_j u_r y_{ro}=1, \qquad (f.4)\\ &\qquad \textstyle\sum_j v_i x_{ij}-\sum_j u_r y_{rj}\geq 0,\ j=1,\ldots,n,\\ &\qquad u_r\geq 0,\ v_i\geq 0,\qquad \forall r,\forall i.\end{aligned}$$

Fig. 1. CCR models for measuring radial efficiency

applications, it is not a good solution for a small sample size or cases with a large number of DMUs with missing values. When a considerable number of DMUs is eliminated from the sample, it may influence the efficiency scores of the remaining DMUs badly. In other words, the efficiency score of the remaining DMUs almost increases; this approach causes some inefficient DMUs to become efficient, and bias results are obtained. And finally, in this approach performance of those DMUs including missing data cannot be evaluated.

3.2 Eliminating outputs and inputs with missing values

Another way to handle missing data is eliminating outputs or inputs variables, including missing data in evaluation. This technique is a common method in statistical science, especially in using SPSS software for data analysis. Again, in the presence of a considerable number of eliminated input/output variables, this approach also yields bias efficiency scores for all units. In this case, the efficiency score of DMUs almost decreases and this approach causes some efficient DMUs to become inefficient. The other problem of this approach, which occurs in rare cases, will happen if each input and output of all DMUs includes at least one missing value; then all the inputs and outputs need to be discarded from the sample, and one may not be able to evaluate the relative performance of the units.

3.3 Assignment value approach

Among the other novel approaches to handling missing data in DEA literature, Kuosmanen (2009) technique is notable, in which missing values are suggested to be replaced by some pre-specified values. Under constant returns to scale (CRS) technology, this approach proposes to replace missing outputs by zero, and missing inputs by a large positive number and then using the classical CCR DEA model for the new data setting.

In the next section, this approach will be discussed in more details and by following its main idea, a new improved technique to handling missing data under DEA framework is proposed.

4. Interval Data Instead of Missing Data

As mentioned earlier, Kuosmanen (2009) suggests the use of dummy entries and replacing missing outputs by zero and missing inputs by sufficiently large positive numbers. This is the first systematic approach to dealing with missing data in the literature. Here, by acknowledging his work, we suggest some points to improve its domain of applicability in the real-world applications. The following discussion shows that Kuosmanen (2009) approach may show some DEA model infeasible.

Taking zero for missing outputs and taking large number for missing inputs is not a proper choice for all production systems. As an instance, let assume all the outputs of DMUO are missed, so by using Kuosmanen method, we need to set $\forall r(r=1,...,s);\ y_{ro}=0$. Now, in the output-oriented envelopment form of the CCR model, there will be a redundant constraint associated with output variables as $\sum_j 0\lambda_j \geq \varphi 0$. This yields to an unbounded objective function value and it is equal to infeasibility of the multiplier form of the CCR model; both are unacceptable. This simple and of course rare case shows that Kuosmanen (2009) approach in handling missing data needs some modifications in order to increase its applicability.

In general, even with preserving feasibility, such assessed dummy values may lead to misleading efficiency scores due to sensitivity problems in DEA as frontier technique based on extreme points. For a set of efficient units with missed inputs/outputs, accepting these values changes the efficient frontier and leads to bias efficiency estimation for all the units even for those with complete data.

To modify Kuosmanen approach and increase its applicability, we suggest using an acceptable range as an interval value for missing data. In real-world applications, there are some estimation techniques to achieve such intervals; for example, by using decision makers (DMs) information on missed values.

To introduce the modified approach, let's assume xi'j and yr'j' are missed in data set for some i', r', j' and j. Based on the decision maker's experience, the following acceptable intervals for these missing data can be presented: $a \leq x_{i'j} \leq b$ and $c \leq y_{r'j'} \leq d$, where the constant bounds a, b, c, and d are real and positive numbers and depending the application could be selected using statistical or experimental techniques.

Now, in constructing the production set based on estimated interval data, the most pessimistic values for the missing items within their intervals are used in our approach, i.e. $x_{i'j}=b, y_{r'j'}=c$ then the classical CCR DEA models for efficiency evaluation of the new data set will be used.

To show the results of this setting formally, let TRD denote the production possibility set determined by "range data" that uses the higher bound of the range for missing inputs and the lower bound of the range for missing outputs. We introduce TRD as follows:

$$T_{RD}=\left\{(x,y)\middle|\begin{cases} x_{i'o}=b\ (\forall i\neq i', x_{io}=\hat{x}_{io}) \\ 0\leq a\leq x_{i'o}\leq b \\ y_{r'o'}=c\ (\forall r\neq r', y_{ro'}=\hat{y}_{ro'}) \\ 0\leq c\leq y_{r'o'}\leq d \\ y_{rj}=\hat{y}_{rj}\ (\forall j\neq o') \\ x_{ij}=\hat{x}_{ij}\ (\forall j\neq o,) \\ (\hat{x}_{ij},\hat{y}_{rj})\in T_{DEA} \end{cases}\right\} \quad (3)$$

TRD, TDMUo, TDEA and TXY are related according to the following theorem.

Theorem 1: Production possibility sets TRD, TDEA, TDMUo, and TXY are nested as:

$$\left.\begin{matrix} T_{DMUo} \\ T_{XY} \end{matrix}\right\} \subseteq T_{RD} \subseteq T_{DEA} \quad (4)$$

Proof: Under constant returns to scale assumption, TDEA, TRD, TDMUo, and TXY could be defined through the following formulas:

Let the r'th output of DMUo is missed.

1) First we show that $T_{RD} \subseteq T_{DEA}$: Due to the formulation of technologies in section 2.1, the only difference between sets T_{RD} and T_{DEA} is in rth output of DMUo. For T_{DEA} this constraint reads as $y_{r'o} \le \sum_{j=1}^{n} \lambda_j y_{r'j}$. Since output r' is missing for DMUo in T_{RD}, the constraint reads as

$$y_{r'o} \le \sum_{j=1}^{o-1} \lambda_j y_{r'j} + \lambda_o \times c + \sum_{j=o+1}^{n} \lambda_j y_{r'j} \qquad (5)$$

$$\le \sum_{j=1}^{o-1} \lambda_j y_{r'j} + \lambda_o \times y_{r'o} + \sum_{j=o+1}^{n} \lambda_j y_{r'j} = \sum_{j=1}^{n} \lambda_j y_{r'j}$$

Thus, the amount of acceptable $y_{r'o}$ in T_{RD} must always be less than or equal to the corresponding value of $y_{r'o}$ in T_{DEA}. Since the two sets are otherwise identical, it is proved that $T_{RD} \subseteq T_{DEA}$.

2) In the second step, it will be shown that $T_{XY} \subseteq T_{RD}$: The constraint for output r' in T_{XY} reads as $y_{r'o} \le \sum_{j=1}^{n} \lambda_j \times 0 = 0$ and the constraint for output r' in T_{RD} reads as

$y_{r'o} \le \sum_{j=1}^{o-1} \lambda_j y_{r'j} + \lambda_o \times c + \sum_{j=o+1}^{n} \lambda_j y_{r'j}$. It is clear that

$$0 \le \sum_{j=1}^{o-1} \lambda_j y_{r'j} + \lambda_o \times c + \sum_{j=o+1}^{n} \lambda_j y_{r'j}$$

So, the acceptable amounts of output r' in set T_{RD} are greater than or equal to the zero value implied by T_{XY}. Since the two sets are otherwise identical, the two sets position are related as $T_{XY} \subseteq T_{RD}$.

3) And finally let show $T_{DMUo} \subseteq T_{RD}$: Comparing T_{DMUo} and T_{RD}, easily can be observed that:

$$y_{r'o} \le \sum_{j=1}^{o-1} \lambda_j y_{r'j} + 0 \times y_{r'o} + \sum_{j=o+1}^{n} \lambda_j y_{r'j}$$

$$\le \sum_{j=1}^{o-1} \lambda_j y_{r'j} + \lambda_o \times c + \sum_{j=o+1}^{n} \lambda_j y_{r'j}$$

Now, $y_{ro} \le \sum_{j=1}^{o-1} \lambda_j y_{rj} + 0 \times y_{ro} + \sum_{j=o+1}^{n} \lambda_j y_{rj}$ for output r ($r \ne r'$), in T_{DMUo} but $y_{ro} \le \sum_{j=1}^{n} \lambda_j y_{rj}$ $(r \ne r')$ in T_{RD} and it is clear that:

$$\sum_{j=1}^{o-1} \lambda_j y_{rj} + 0 \times y_{ro} + \sum_{j=o+1}^{n} \lambda_j y_{rj} \le \sum_{r=1}^{s} \lambda_j y_{rj} \quad (r \ne r')$$

Thus, the amount of acceptable $y_{r'o}$ in T_{DMUo} must always be less than or equal to the corresponding value of $y_{r'o}$ in T_{RD} for any output. So, it is proved that $T_{DMUo} \subseteq T_{RD}$.

The similar proof can be presented for missing inputs or for the case that missing inputs and outputs are existed.

5. Numerical Examples

In this part of the paper, through three numerical examples, a comparison on the results of the above-mentioned approaches in dealing with missing data will be made, and it will be shown that our proposed approach is more applicable than the other ones. The same data set as example 1 was analyzed in Kuosmanen (2009) paper. Data set of examples 2 and 3 are similar to the data set of some examples in research of Kuosmanen (2009).

For simplicity of presentation, let θ_I be used to show the efficiency score of units when all the data are available; θ_{II} for the efficiency score of units when the DMUs with missing values are omitted from the observation; θ_{III} shows the efficiency score of units when the missing outputs or missing inputs are omitted from the data set; θ_{IV} also shows the efficiency score of DMUs obtained by Kuosmanen (2009) approach and finally θ_V denotes the efficiency score of DMUs, which are obtained by our interval approach. The data, which are in parentheses, are supposed be the exact values of the missing inputs or outputs.

Example 1. Table 1 shows data for 5 hypothetical DMUs with one input X, and two outputs Y. The three last columns show an acceptable range of the missing items.

Table 1
Five units including missing output values

DMU	X	Y1	Y2	Missing X	Missing Y1	Missing Y2
A	1	15	45	-	-	-
B	1	(20)	60	-	[15,30]	-
C	1	35	40	-	-	-
D	1	(45)	30	-	[30,50]	-
E	1	50	10	-	-	-

Table 2 reports efficiency score of the units computed by the five approaches by using output orientation of the CCR model.

Table 2
Efficiency scores for five units

DMU	θ_I	θ_{II}	θ_{III}	θ_{IV}	θ_V
A	0.75	1	0.75	0.89	0.8
B	1	-	1	1	1
C	0.98	1	0.67	1	1
D	1	-	0.5	0.5	0.82
E	1	1	0.17	1	1

Based on the results, when all data are known, the first approach shows DMUs B, D, and E are efficient. The

results in the second column show the strategy of omitting units, including missing data from the observation is beneficial for remaining units, for example, DMU A, which is not efficient in the original data set, becomes efficient. In contrast with the first approach, the results in the third column show when the output variable concluding missing value (Y2) is omitted, some of the efficient units, like D and E become inefficient in the new data set. It is caused by neglecting their good output performance in comparison with the other units. The fourth column shows efficiency scores computed by the Kuosmanen (2009) approach. In comparison with the second column, this approach suggests efficiency scores for all the units, and the results are more accurate in comparison with the previous approaches. The results of the new interval approach are reported in the last column of Table 3. The results show that the new approach gives a more accurate efficiency score in comparison with the original data results, as well as the Kuosmanen (2009) approach. However, there is a notable difference; DMU D, which is an efficient unit regarding the complete data set, gets a better rank in the interval approach compared to its score in Kuosmanen (2009) approach.

The differences could be even more in real data set when more accurate approximations are used in the form of intervals, in which the unknown missing values are likely to belong, and the computed results of the modified approach will be more reliable.

Example 2. Table 3 shows data for five hypothetical DMUs with two inputs X, and two outputs Y. The four last columns are used to show a selected range of missing inputs and outputs.

Table 3
Data for five units

DMU	X1	X2	Y1	Y2
A	20	15	45	30
B	(35)	20	60	20
C	30	35	(40)	15
D	(40)	45	30	50
E	25	50	(10)	45
DMU	Missing X1	Missing X2	Missing Y1	Missing Y2
A	-	-	-	-
B	[25,40]	-	-	-
C	-	-	[30,45]	-
D	[30,50]	-	-	-
E	-	-	[5,15]	-

The computed efficiency scores for this data setting in the five models are summarized in the table 4.

Table 4
Efficiency scores for five units

DMU	θ_I	θ_{II}	θ_{III}	θ_{IV}	θ_V
A	1	1	1	1	1
B	1	-	0.5	1	1
C	0.59	-	0.21	0.31	0.44
D	0.79	-	0.56	0.56	0.65
E	1	-	0.45	1	1

As it can be seen, scores in the last column computed using the new approach, gives the best approximation of the original efficiency scores among the other techniques.

Example 3. As the last illustration example, consider the table 5, which shows data for five hypothetical DMUs with one input X, and one output Y. The two last columns of the table shows suggested ranges for missing data.

Table 5
the data of five units

DMU	X1	Y1	Missing X1	Missing Y1
A	20	45	-	-
B	(35)	60	[25,40]	-
C	30	(40)	-	[30,45]
D	(40)	30	[30,50]	-
E	25	(10)	-	[5,15]

The computed efficiency scores using five approaches are included in Table 6.

Table 6
Efficiency scores for five units

DMU	θ_I	θ_{II}	θ_{III}	θ_{IV}	θ_V
A	1	1	-	1	1
B	0.76	-	-	0.27	0.67
C	0.59	-	-	Infeasible	0.44
D	0.33	-	-	0.13	0.27
E	0.18	-	-	Infeasible	0.09

As the results show, two approaches are failed in computing efficiency scores for the units, including missing data. In addition, the method proposed by Kuosmanen (2009) cannot suggest an efficiency score for units C and D, and the new approach is the only technique, which finds feasible efficiency scores near to the original values.

6. Conclusion

One of the basic assumptions of the original DEA is that inputs and outputs are known on a ratio scale for all units. However, in many DEA applications, it is practically impossible to collect complete data sets, and missing values in some input/output variables are inevitable. In DEA literature, there have been few approaches, which are introduced for dealing with missing data.

Through some simple numerical examples, it has been shown that the previous works like eliminating DMUs with missing values or eliminating missing inputs or

outputs may destroy the true classification of the units into efficient and inefficient ones. Also, these approaches may fail in measuring the efficiency scores of some units. One attempt in using missing data was done by Kuosmanen (2009), which may make infeasible solution for some DMUs. The current paper follows and extends the way proposed by Kuosmanen (2009) by assigning appropriate values as an acceptable interval for the missing data applying decision maker's information. Then the worst bounds of the suggestion interval for variables with missing data are used to estimate the performance of DMUs. With some numerical examples, it was shown that the new approach is more applicable than the others in the existing literature.

To address the many practical problems encountered in settings such as these, we suggest that more extensive research be done on evaluating economical efficiencies such as cost efficiency or profit efficiency for units with missing prices.

References

Azizi H. (2013), "A note on data envelopment analysis with missing values: an interval DEA approach", The International Journal of Advanced Manufacturing Technology, 66(9-12), 1817-1823.

Banker, RD, Charnes, A., Cooper, WW. (1984), "Models for Estimation of Technical and Scale Inefficiencies in Data Envelopment Analysis", Management Sience, 30, 1078-1092.

Charnes A, Cooper WW. (1962), "Programming with linear fractional functional", Naval Research Logistics Quarterly, 9, 181-185.

Charnes A, Cooper WW, Rhodes E. (1978), "Measuring the efficiency of decision making units", European journal of operational research, 2(4), 429 - 444.

Cooper WW, Seiford LM, Tone K. (2000), Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-Solver Software, Kluwer Academic Publishers, Boston.

Farrell MJ. (1957), "The Measurement of Productive Efficiency", Journal of Royal Statistical Society, 120(3), 253-281.

Kao C, Liu ST, (2000), "Data envelopment analysis with missing data: An application to University Libraries in Taiwan", Journal of the Operational Research Society, 51 (8), 897–905.

Kuosmanen, T. (2009), Data envelopment analysis with missing data, Journal of the Operational Research Society, 60, 1767-1774.

Neal, PVO, Ozcan, YA, Yanqiang, M. (2002), "Benchmarking mechanical ventilation services in teaching hospitals", Journal of Medical Systems. 26(3), 227–240.

Smirlis YG, Maragos EK, Despotis DK. (2006), "Data envelopment analysis with missing values: An interval DEA approach". Applied Mathematics and Computation, 177(1), 1-10.

Zha Y, Song A, Xu Ch, Yang H. (2013), "Dealing with missing data based on data envelopment analysis and halo effect", Applied Mathematical Modelling, 37(9), 6135–6145.