DATA ENVELOPMENT ANALYSIS MODELS WITH MISSING VALUES

[1]JOSEF JABLONSKY, [2]LUBOS MAREK

University of Economics, Prague, Czech Republic
E-mail: [1]jablon@vse.cz, [2]marek@vse.cz

Abstract - Data envelopment analysis models are generally used tools for efficiency and performance evaluation of a set of homogenous decision-making unit. It is based on a comparison of the inputs that the units spend for the production and the outputs produced relative to the other units of the set. The paper deals with the case that some of the inputs or output values are missing – this can be given by their non-availability or by the non-homogeneity of the data set. The paper discusses possible ways how to solve data envelopment analysis models with missing values. A numerical example illustrates possible approaches and their advantages and drawbacks are analyzed.

Index Terms - Data envelopment analysis, missing data, non-homogeneity, efficiency

I. INTRODUCTION

Data envelopment analysis (DEA) was introduced in 1978 by Charnes et al. [1] as a generalization of Farrell's concept of efficiency for multiple inputs and multiple outputs, published in [2]. DEA models measure the efficiency of transformation of multiple inputs into multiple outputs of a decision-making unit (DMU) relatively with respect to other DMUs of a homogeneous set of DMUs. Let the set of DMUs has n elements DMU_1 to DMU_n. The units spend m inputs for the production of r outputs.Let x_{ij}, i = 1, …, n, j = 1, …, m be the j-th input of the DMU_i, and y_{ik}, i = 1, …, n, k = 1, …, r be the k-th output of the same unit. The efficiency score of the DMUq can be expressed as

$$\theta_q = \frac{\sum_{k=1}^{r} u_k y_{qk}}{\sum_{j=1}^{m} v_j x_{qj}}$$

where u_k, k = 1, …, r, and v_j, j = 1, …, m, are the weights of the outputs and inputs respectively. Thus, the efficiency score is defined as the weighted sum of outputs divided by the weighted sum of inputs where weights are to be determined as the variables of the model in order to maximize the ratio (1).

DEA models assume that all data (inputs and outputs) are known and they are non-negative. Another assumption consists in the homogeneity of the set of DMUs. The problem is that very often some values of inputs and/or outputs are not available, they are missing. It can be caused by several reasons. At first, the data for some DMUs and some inputs/outputs are not available to decision makers. Another reason for missing data can be caused by the non-homogeneity of the data set. It will be discussed further in the paper. It is focused on a discussion about possible ways how to deal with missing data in DEA models.

The paper is organized as follows. Next section contains the formulation of traditional DEA models. Section III presents several possible approaches for dealing with missing data. Section IV illustrates their using on a numerical example. The final section concludes the paper and discusses possible directions of future research.

II. DATA ENVELOPMENT ANALYSIS

Traditional DEA model introduced in [1] maximizes the efficiency score of the unit under evaluation (1) under the assumption that efficiency scores of all other units of the set do not exceed 1, and the weights are strictly positive. Its mathematical formulation is as follows:

Maximize

$$\theta_q = \frac{\sum_{k=1}^{r} u_k y_{qk}}{\sum_{j=1}^{m} v_j x_{qj}} \quad (1)$$

subject to (2)

$$\frac{\sum_{k=1}^{r} u_k y_{qk}}{\sum_{j=1}^{m} v_j x_{qj}} \le 1, \qquad i = 1,...,n,$$

$$v_j \ge \varepsilon, \quad j = 1,...,m,$$

$$u_k \ge \varepsilon, \quad k = 1,...,r,$$

where ε is an infinitesimal constant that ensures strict positivity of the weights.

Model (2) is denoted usually as CCR (Charnes, Cooper, Rhodes) model. It is not linear in the objective function but can be moved to a linear program using Charnes and Cooper transformationeasily.The linearized version of the so-called input-oriented CCR model is formulated as follows:

Maximize

$$\theta_q = \sum_{k=1}^{r} u_k y_{qk}$$

subject to (3)

$$\sum_{k=1}^{r} u_k y_{qk} - \sum_{j=1}^{m} v_j x_{qj} \le 0, \qquad i = 1,...,n,$$

$$\sum_{j=1}^{m} v_j x_{qj} = 1,$$

$$v_j \ge \varepsilon, \quad j = 1,...,m,$$

$$u_k \ge \varepsilon, \quad k = 1,...,r,$$

Models (2) and (3) create the efficient frontier with the assumption of constant returns to scale. They have been generalized in [3] for the case of variable, and more generally non-constant returns to scale technologies. Model (3) has (n+m+r+1) constraints and (m+r) variables. In order to derive efficiency scores of all units, model (3) has to be solved n times (for each unit of the set). From the computational point of view, it might be better to work with the dual model of model (3):

Minimize

$$\theta_q - \varepsilon \left(\sum_{j=1}^{m} s_j^- + \sum_{k=1}^{r} s_k^+ \right)$$

subject to

$$\sum_{i=1}^{n} x_{ij} \lambda_i + s_j^- = \theta_q x_{qj}, \quad j = 1,...,m,$$

$$\sum_{i=1}^{n} y_{ik} \lambda_i - s_k^+ = y_{qk}, \quad k = 1,...,r,$$

$$s_k^+ \ge 0, \qquad k = 1,...,r,$$

$$s_j^- \ge 0, \qquad j = 1,...,m,$$

$$\lambda_i \ge 0, \qquad i = 1,...,n.$$

The DMUs with their efficiency scores $\theta_q = 1$ are efficient units (fully or weakly). The inefficient units have their efficiency scores $\theta_q < 1$. More about the theory of DEA models can be found in [4], [5], or [6].

III. MISSING DATA IN DEA

Let us suppose that the set of DMUs is divided into two disjoint subsets U_1 and U_2. The first subset contains the units with all data available. The second one has some missing data. We can consider the case that second group of units does not produce a specific output that is produced by the first group of DMUs. In order to evaluate the efficiency of the units of both groups simultaneously the following approaches are proposed by various authors:

- Simply replace the missing values by zeroes, and then a traditional DEA model can be applied.
- To replace the missing values with optimistic, pessimistic or most likely estimates. They can be derived by comparison of the inputs and outputs of the first group of units where all data are available.
- To apply a procedure that was proposed in Cook et al. [7]. Their approach is based on an idea to reduce the inputs of the units from the first group by a certain percentage α because they suppose that this amount of inputs is used for the production of the outputs missing in the second group of units. The procedure consists of two steps:
  - In the first step, all inputs of the first group are reduced by α %. Only outputs produced by both groups of units are considered and a traditional DEA model is solved for all units of the set. The efficiency scores obtained in this step for the units in U_2 are their final scores.
  - The second step considers the units in U_1 only. This step uses α % of all inputs and the remaining part of outputs that are not produced by the units in U_2. The final efficiency scores for the units in U_1 are calculated as the weighted averages of their scores derived in both steps.

In this paper we offer a procedure that is based on replacing the missing values by interval values, i.e. the missing values will be replaced by $\left\langle y_{ik}^L, y_{ik}^U \right\rangle, i \in U_2$, where y_{ik}^L is the lower bound for the k-th output and y_{ik}^U is the upper bound for the k-th output. These values can be estimated either by the decision-maker or derived from the known inputs and outputs in the first group of units. The lower and upper bounds for the efficiency scores of the DMU_q in U_2 can be obtained by solving the following couple of optimization problems that are derived from the approach proposed in Despotis and Smirlis [8]:

Minimize

$$\theta_q^L$$

Subject to (5)

$$\sum_{i=1}^{n} \left(\lambda_i x_{ij} \right) + s_j^- = \theta_q^L x_{qj}, \qquad j = 1,...,m,$$

$$\sum_{i=1, i \ne q}^{n} \left(\lambda_i y_{ik}^U \right) + \lambda_q y_{qk}^L - s_k^+ = y_{qk}^L, \qquad k = 1,...,r,$$

$$s_k^+ \ge 0, \qquad k = 1,...,r,$$

$$s_j^- \ge 0, \qquad j = 1,...,m,$$

$$\lambda_i \ge 0, \qquad i = 1,...,n.$$

and

Minimize

$$\theta_q^U$$

Subjectto (6)

$$\sum_{i=1}^{n} \left(\lambda_i x_{ij} \right) + s_j^- = \theta_q^U x_{qj}, \qquad j = 1,...,m,$$

$$\sum_{i=1, i \ne q}^{n} \left(\lambda_i y_{ik}^L \right) + \lambda_q y_{qk}^U - s_k^+ = y_{qk}^U, \qquad k = 1,...,r,$$

$$s_k^+ \ge 0, \qquad k = 1,...,r,$$

$$s_j^- \ge 0, \qquad j = 1,...,m,$$

$$\lambda_i \ge 0, \qquad i = 1,...,n.$$

Model (5) considers lower bound for outputs for the

unit under evaluation (DMU_q), i.e. the worse value for the outputs, and upper bound for outputs for all other units. The results of the lower bound for the efficiency score θ_q^L. On the contrary, model (6) uses the best values for the outputs for the unit under the evaluation and the worse values for all other units. This model derives the upper bound for the efficiency score θ_q^U .Please note that for the known output values the lower and upper bounds are identical. According to the lower and upper bounds the units can be split into the following three groups:

- $\theta_q^L = 1$ - The units always efficient.
- $\theta_q^U = 1$ and $\theta_q^L < 1$ - the units efficient only in case of a suitable setting of missing outputs.
- $\theta_q^U < 1$ - the units never efficient.

All units of the set can be ranked according to the lower and/or upper bounds of efficiency scores. The upper bound is often equal to the maximum value (1) which does not allow ranking of the units. In order to distinguish among them some of the super-efficiency models can be applied. The first model of this category was introduced by Andersen and Petersen [9] in 1993. Jablonsky [10] proposed a super-efficiency model that improves some its drawbacks.

Table 1: Data set

DMU	X1	X2	Y1	Y2
DMU 1	54284	1934	91	1794
DMU 2	48879	454	26	744
DMU 3	96830	1269	49	800
DMU 4	54595	790	22	470
DMU 5	85779	1391	21	1281
DMU 6	46512	1122	39	1481
DMU 7	53834	1798	90	865
DMU 8	70780	536	25	774
DMU 9	72830	845	20	1287
DMU 10	53207	1501	91	1624
DMU 11	90563	487	50	585
DMU 12	53271	641	85	1475
DMU 13	47361	1685	31	1722
DMU 14	70816	861	77	1201
DMU 15	51629	872	25	
DMU 16	23814	240	28	
DMU 17	21179	609	22	
DMU 18	32474	732	27	
DMU 19	47789	974	52	
DMU 20	53584	263	28	
DMU 21	49254	223	33	
DMU 22	30295	682	46	
DMU 23	34424	259	58	

Unfortunately, the ranges obtained by the approach presented above may be quite wide very often which is not an acceptable conclusion for decision-makers. One of the possibilities how to overcome this drawback is to use simulation procedure that allows estimating the probability of being on certain rank positions for all DMUs. Instead of interval values, it is possible to define a probability distribution and randomly generate the values of outputs according to this assumption. In particular simulation trials, the simulation approach computes the efficiency scores and ranking of all units. By repeating the trials, it is possible to estimate the probability of ranking of the units. Of course, this approach is much computationally demanding than the optimization approach.More information about this approach can be found in Jablonsky [11].

IV. A NUMERICAL EXAMPLE

All approaches will be illustrated on a numerical example in this section of the paper. The data set is a modified and simplified version of the set contained in [12]. There are 23 DMUs in the data set. The subset U_1 consists of 14 units that spend two inputs and produce two outputs. The second subset has 9 units that spend the same two inputs but produce only one output which is identical with the first output of the units in U_1. The complete data set is included in Table 1.

At first, we will derive the lower and upper bounds for the intervals of Y2 output for the units in U_2. From the data for units in U_1, we can see that the lowest ratio Y2/X2 (units of the second output per one unit of the second input) equals 0.4811 for DMU_7, and the highest ratio is 2.3011 for DMU_{12}. Using these coefficients for the units in U_2 we have the lower and upper bounds with respect to the ratio Y2/X2. Similarly we can derive the bounds based on the ratio Y2/X1. Both sets of bounds are included in Table 2.

Table 2: Lower and upper bounds for output Y2

	Y2/X2		Y2/X1	
DMU	Lower	Upper	Lower	Upper
DMU 15	420	2007	334	1877
DMU 16	115	552	154	866
DMU 17	293	1401	137	770
DMU 18	352	1684	210	1181
DMU 19	469	2241	309	1738
DMU 20	127	605	346	1948
DMU 21	107	513	318	1791
DMU 22	328	1569	196	1101
DMU 23	125	596	222	1252

Table 3 contains the results of numerical experiments. The following results are included:

- Column (1) – DMUs.
- Column (2) – Model 1: Efficiency (or super-efficiency for efficient units) scores for the data set where the missing values are replaced by zeros.
- Column (3) –Model 2: Efficiency (or super-efficiency for efficient units) scores for the data set where the missing values are replaced by

average values calculated from the relation Y2/X2 for all units in U_1 which is 1.0515. E.g. for DMU_{15} its missing value is calculated as 872·1.0515 = 917.

- Columns (4) and (5) – Model3: Lower and upper bounds of efficiency (super-efficiency) scores obtained using models (5) and (6) with lower and upper bounds for the missing values taken from the first two columns of Table 2.
- Columns (6) and (7) – Model 4: Lower and upper bounds of efficiency (super-efficiency) scores obtained using models (5) and (6) with lower and upper bounds for the missing values taken from the last two columns of Table 2.

Table 3: Efficiency (super-efficiency) scores

(1)	(2)	(3)	(4)	(5)	(6)	(7)
DMU 1	1.058	1.059	0.994	1.058	0.988	1.058
DMU 2	0.712	0.712	0.702	0.712	0.419	0.712
DMU 3	0.314	0.314	0.307	0.314	0.299	0.314
DMU 4	0.301	0.301	0.257	0.301	0.239	0.297
DMU 5	0.511	0.511	0.394	0.511	0.411	0.500
DMU 6	0.991	0.991	0.581	0.991	0.875	0.938
DMU 7	0.977	0.978	0.977	0.977	0.977	0.977
DMU 8	0.627	0.628	0.618	0.628	0.301	0.628
DMU 9	0.662	0.662	0.651	0.662	0.486	0.662
DMU 10	1.032	1.030	1.022	1.031	1.015	1.030
DMU 11	0.671	0.633	0.522	0.657	0.458	0.642
DMU 12	1.618	1.618	1.000	1.605	0.944	1.618
DMU 13	1.100	1.100	0.574	1.100	1.000	1.013
DMU 14	0.676	0.674	0.659	0.675	0.643	0.675
DMU 15	0.285	0.603	0.291	1.319	0.285	1.233
DMU 16	0.697	0.708	0.697	1.000	0.697	1.568
DMU 17	0.607	0.912	0.607	1.954	0.607	1.078
DMU 18	0.488	0.751	0.488	1.643	0.488	1.152
DMU 19	0.640	0.716	0.640	1.524	0.640	1.182
DMU 20	0.475	0.610	0.475	1.000	0.475	3.219
DMU 21	0.661	0.748	0.661	1.062	0.661	3.490
DMU 22	0.891	0.898	0.891	1.642	0.891	1.167
DMU 23	1.621	1.621	1.621	1.621	1.621	2.101

Table 3 contains many numbers that can be explained from various points of view:

- The results of Model 1 and Model 2 show a very close similarity of the efficiency scores for the units in U_1. It is not surprising that the efficiency scores for the units in U_2 obtained by Model 2 are higher or equal than those given by Model 1.
- Lower bounds of efficiency scores obtained by both Models 3 and 4 for the units in U_2 are identical, i.e. the scores given by the first two models are their theoretically minimum values.
- Application of model (6) leads to the conclusion that all units in U_2 are efficient, i.e. maximum super-efficiency score is greater than 1. In several cases, especially in the second set of intervals, the super-efficiency scores are very high (DMU_{20} and DMU_{21}) which probably does not correspond to the reality. It is given by very wide intervals for these units compared to their inputs.
- The scores in columns (4) to (7) for the units in U_1 are not theoretically lower and upper bounds of efficiency scores but just minimum and maximum values coming from the application of models (5) and (6).

CONCLUSIONS

The paper discussed several possible ways how to deal with missing values in applications of DEA models. Replacing the missing values by their interval estimates is a good idea, but the model returns just lower and upper bounds for the efficiency scores. Some of them are hard to explain. A simulation approach is more computationally demanding, but its results have much higher information value. Future research will be focused on an extension and

generalization of the proposed procedures and their comparison with simulation approaches.

REFERENCES

[1] A. Charnes, W. W. Cooper, and E. Rhodes, "Measuring the efficiency of decision-making units," European Journal of Operational Research, vol. 2, no. 6, pp. 429–444, 1978.

[2] M. J. Farrell, "The measurement of productive efficiency", Journal of the Royal Statistical Society,vol. 120, pp.253-290, 1957.

[3] R. D. Banker, A. Charnes, and W. W. Cooper. "Some models for estimating technical and scale inefficiencies in data envelopment analysis," Management Science, vol. 30, no. 9, pp. 1078–1092, 1984.

[4] W. W. Cooper, L. M. Seiford, and K. Tone, Data envelopment analysis: A comprehensive text with models, applications, references and DEA-solver software, Berlin, Springer, 2006.

[5] J. Zhu, Quantitative models for performance evaluation and benchmarking: Data envelopment analysis with spreadsheets. Boston, Springer, 2009.

[6] M. Dlouhy, J. Jablonsky, and P. Zykova, Data envelopment analysis (in Czech). Praha, Professional Publishing, 2018.

[7] W. D. Cook, J. Harrison, R. Imanirad, P. Rouse, and J. Zhu, "Data envelopment analysis with nonhomogeneous DMUs" Operations Research, vol. 61, no. 3, pp. 666–676, 2013.

[8] D. Despotis and Y. Smirlis, "Data envelopment analysis with imprecise data," European Journal of Operational Research, vol. 140, no. 1, pp. 24-36, 2002.

[9] P. Andersen, and N. C. Petersen, "A procedure for ranking efficient units in data envelopment analysis", Management Science, vol. 39, pp. 1261 –1264, 1993.

[10] J. Jablonsky, "Multicriteria approaches for ranking of efficient units in DEA models", Central European Journal of Operations Research, vol. 20, pp. 435–449, 2012.

[11] J. Jablonsky, "Two-stage data envelopment analysis model with internal inputs and outputs", International Journal of Trade, Economic and Finance, vol. 4, no. 1, pp. 55-59, 2013.

[12] Y. Chen, L. Wu, and B. Lu,"Data envelopment analysis procedure with two non-homogeneous DMU groups", Journal of Systems Engineering and Electronics, vol. 29, no. 4, pp. 780-788, 2018.

★★★