No. 08-01
2008

CONSEQUENCES OF OMITTING RELEVANT INPUTS ON THE QUALITY OF THE DATA ENVELOPMENT ANALYSIS UNDER DIFFERENT INPUT CORRELATION STRUCTURE.

Andrés Ramírez Hassan

Documentos de trabajo
Economía y Finanzas
Centro de Investigaciones Económicas y Financieras (CIEF)

UNIVERSIDAD EAFIT
Abierta al mundo

Foto: Juan Carlos Ramírez

Consequences of omitting relevant inputs on the quality of the data envelopment analysis under different input correlation structures

Andrés Ramírez Hassan[∞]

Abstract: *This paper establishes the consequences of a wrong specification on the quality of the data envelopment analysis. Specifically, the case of omitting a relevant variable in the input oriented problem is analyzed when there are different correlation structures between the inputs. It is established that the correlation matrix gives relevant information about the homogeneity of the decision making units and the intensity of inputs used in the production process. The methodology is based on a series of Monte Carlo simulations and the quality of the data envelopment analysis is measured as the difference between the true efficiency and the efficiency calculated. It is found that omitting relevant inputs causes inconsistency, and this problem is worse when there is a negative correlation structure.*

Keywords: *Efficiency, Data Envelopment Analysis, Monte Carlo Simulation, Input Correlation Structure.*

[∞] Department of Economics, EAFIT University, Medellín, Colombia. Email: aramir21@eafit.edu.co.

Introduction

Data Envelopment Analysis (DEA) is a nonparametric tool based on mathematical programming that is utilized to calculate the relative efficiency of a set of Decision Making Units (DMUs) which operate in homogeneous conditions. The technique has been frequently used since it was introduced by Charnes, A., Cooper, W. and Rodhes, E. in 1978.

DEA estimates the minimum combination of inputs for producing some given outputs (*input oriented*) or the maximum combination of outputs that can be achieved with some inputs (*output oriented*). In theory, this means that the technique constructs the relative isoquant or the relative production frontier, respectively. This is done based on the behavior observed empirically. Basically, the methodology generates linear combinations between the efficient DMUs in order to create virtual producers that are the reference points for calculating the inefficiencies of all the DMUs. This is the reason for saying that the efficiencies estimated are relative.

This tool computes a scalar measurement of efficiency and determines the efficient levels of inputs and outputs for the DMU under valuation. This kind of information is useful for determining the critical points that can be modified by the managers in order to achieve better outcomes. It is necessary to say that if there are uncontrollable factors that affect the production process, this phenomenon should be taken into consideration given that this will modify the final efficiency ranking.

The technique has many advantages: first of all, the researcher does not have to assume a specific production function or cost function, this implies a recognition of the differences in the production process between DMUs. Second, the optimization process that is needed in order to evaluate performance is done for each DMU, then the parameters estimated belong to each DMU. Third, the methodology permits the combining of different measurement units. Fourth, the researcher can handle simultaneously DMUs that produce many outputs and use different inputs. Fifth, although the first works did not discriminate between technical efficiency and scale efficiency (Charnes, Cooper and Rodhes, 1978, 1979 and 1981), a later version established a model that permits the handling of variable returns to scale (Banker, Charnes and Cooper, 1984). Also, a new model postulated by Färe and Grosskoph (1985), which introduces the market prices, permits the evaluating of allocative efficiency.

On the other hand, the tool has some disadvantages that have to be mentioned. Given that the tool constructs the virtual producers as a linear combination of efficient DMUs to evaluate the relative efficiency, a DMU that uses an unrealistic combination of factors can be judged as efficient due to there being no other DMUs that use such extreme combinations. However, there is a possible solution to this situation, this consists of imposing some restrictions on the estimated weights in order to establish an assurance region (Thompson, Langemeier, Lee and Thrall, 1990; Charnes, Cooper, Huang and Sun, 1990), with this procedure the unrealistic combinations are judged inefficient. This mechanism requires an interaction between the researcher and the administrator, who know the technical problems faced by the DMUs. This should be done in order to impose good

restrictions on the parameters; however, subjective arguments can change the outcomes drastically. Another way to handle the problem of extreme input combinations is to introduce the allocative efficiency criterion. This has the advantage of utilizing market prices to judge the performance of the DMUs. Second, given that DEA is a nonparametric tool, the measurement errors can significantly affect the outcomes, because the analysis is established in comparative conditions; special attention has to be put on efficient units, due to measurement errors cause bias. Third, it is not possible to do statistical inference over individual weights, because DEA does not yet have a well developed statistical foundation; however, Banker (1993) showed that DEA estimators of the best practice monotone increasing and concave production function are also maximum likelihood estimators, while the best practice estimator is biased in finite sample size, the bias approaches zero for large samples. This result is the basis for statistical inference for groups inside the sample. Also, given that DEA does not incorporate random effects, this can not discriminate technical inefficiency and stochastic shocks that affect the production process. Moreover, DEA analysis has outcomes less stable than the outcomes obtained in regression analysis, because the comparative analysis for each DMU is done in a small group from the sample. Another criticism of the technique is that the efficiency calculated is relative and the process has a slow convergence to the global optimum.

Although DEA is a flexible and useful tool for evaluating performance of DMUs, the outcomes calculated depend on the true efficiencies, but also, there are other factors affecting the outcomes. Special attention has to be put on the model specification in order to achieve robustness (Pedraja, Salinas and Smith, 1999). Moreover, the data used has to be carefully checked for eliminating possible measurement errors. It is necessary to analyze simultaneously enough DMUs (minimum three DMUs for each factor; Bancker, 1989). Finally, the variables that will be included in the analysis have to be studied, specifically their characteristics and correlations.

If regression analysis is used to determine efficiency, there are basically two model specification problems; an incorrect design matrix and a wrong functional form. These problems cause serious consequences over the estimated parameters' properties. On the other hand, when DEA is utilized to analyze efficiency, it is not necessary to specify a functional form; however, there are other specification problems, an incorrect input or output structure, not taking into account the effect of uncontrollable variables or using a wrong orientation problem. Obviously, these specification mistakes affect the efficiency calculated by DEA.

The objective of this paper is to show the consequences of a wrong specification on the quality of DEA; specifically, the problem of omitting relevant variables into the efficiency analysis is studied when different correlation input structures are supposed.

This paper is organized as follows: the first section shows the basic mathematical foundation of DEA, the second section establishes the basis for the Monte Carlo simulations exercises, the third section shows the results, and finally, some concluding remarks are enunciated.

I. Basic mathematical foundations of DEA

The model that was initially proposed by Charnes et al. (1978) is given in (1). This model is designed to evaluate the relative efficiency of the DMU_0. This one produces *s* outputs ($y_{r0} > 0$) and utilizes *m* inputs ($x_{i0} > 0$) (Charnes, Cooper and Thrall, 1991, showed how the positive restrictions can be relaxed).

The optimization problem consists of finding the optimum weights, u_r^* and v_i^*, that maximize h_0 subject to the restrictions. The first restriction establishes that the process is stopped when some DMUs reach $h = 1$, then the DMU_0's maximum efficiency score will be $h_0^* \leq 1$. The DMUs that reach this limit provide the basis to build the virtual producer which is necessary to evaluate the relative performance of the DMU_0. Moreover, the restrictions implicate that all the weights have to be positive and the isotonicity property has to be fulfilled, this means that an increase in any input should result in some output increase and not a decrease in any output.

The ε represents a non-archimedean constant which is smaller than any positive valued real number. The process has to be done *n* times in order to establish the efficiency level for each DMU.

$$\underset{u_r, v_i}{Max} \quad h_0 = \frac{\sum_{r=1}^{s} u_r y_{r0}}{\sum_{i=1}^{m} v_i x_{i0}}$$

s.t.

$$\frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{j=1}^{m} v_i x_{ij}} \leq 1, \quad j = 1,2,...,n \qquad (1)$$

$$\frac{u_r}{\sum_{i=1}^{m} v_i x_{i0}} > \varepsilon, \quad r = 1,2,...,s$$

$$\frac{v_i}{\sum_{i=1}^{m} v_i x_{i0}} > \varepsilon, \quad i = 1,2,...,m$$

$$\varepsilon > 0$$

This model is common to find in the following way (see (2)). This representation is based on the theory of fractional programming.

$$\underset{u_r, v_i}{Max} \quad h_0 = \sum_{r=1}^{s} u_r y_{r0}$$

s.a

$$\sum_{r=1}^{s} u_r y_{rj} - \sum_{j=1}^{m} v_i x_{ij} \leq 0 \qquad (2)$$

$$\sum_{i=1}^{m} v_i x_{i0} = 1$$

$$u_r, v_i \geq \varepsilon > 0$$

This primal problem has *n + 1 + s + m* restrictions, while the number of parameters that have to be calculated is *s + m*. This implies that the dual problem will have *s + m* restrictions and *n + 1 + s + m* parameters. Then the dual problem offers some advantages because it has less restrictions. The dual version of the problem can be seen in (3). In this context θ, λ_j, s_i^- and s_r^+ are parameters that have to be calculated, s_i^- and s_r^+ are slacks that determine the optimum level of inputs and outputs that would have to utilize and produce the DMU.

$$\underset{\theta, \lambda_j, s_r^+, s_i^-}{Mim} \quad \theta - \varepsilon \left[\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+ \right]$$

s.a

$$\theta x_{i0} - \sum_{j=1}^{n} x_{ij} \lambda_j - s_i^- = 0 \qquad (3)$$

$$y_{r0} = \sum_{j=1}^{n} y_{rj} \lambda_j - s_r^+$$

$$\lambda_j, s_i^-, s_r^+ \geq 0$$

θ *unrestricted*

If it is introduced the additional restriction $\sum_{j=1}^{n} \lambda_j = 1$ in (3), the solution found has in consideration technical efficiency and scale efficiency for each DMU (Banker et al.,1984).

II. Basis for the Monte Carlo simulations exercises

In order to analyze the problem of omitting relevant variables on the quality of DEA a series of Monte Carlo exercises were done. Specifically, it is supposed a Cobb–Douglas production function with two arguments.

$$y_j = x_{1j}^{0,5} x_{2j}^{0,5}, \quad j = 1,2,...,1000 \qquad (4)$$

Given that DEA's outcome can be influenced by the distribution of the true efficiencies, it is assumed that all the DMUs are efficient. Moreover, there are only three factors (two inputs and one output) and one thousand DMUs, this is done in order to have many degrees of freedom and achieve convergence to true efficiency (consistency). Finally, if there is a correct specification in the analysis, the efficiencies calculated equal the true efficiencies.

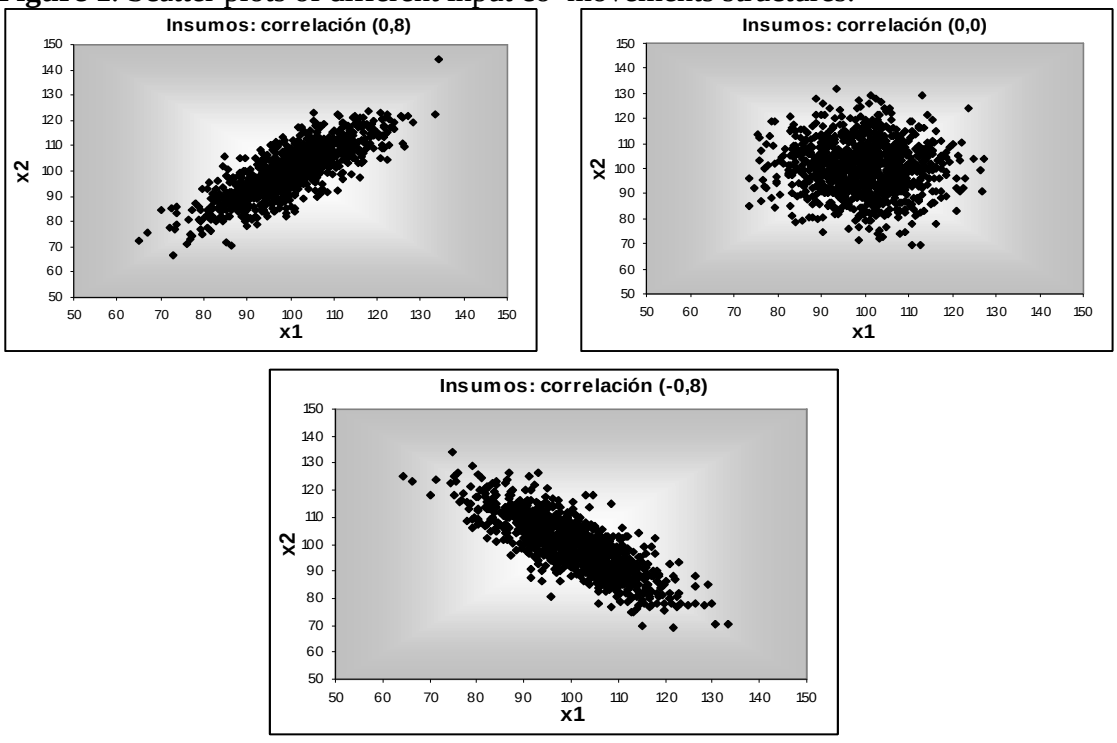
The inputs were generated from a multi–normal distribution with mean one hundred and standard deviation ten. It was supposed three correlations structures between the inputs: independent co–movements (zero correlation), high positive co–movements (0.8 correlation) and high negative co–movements (-0.8 correlation). For each correlation structure five hundred simulations were generated. A total of 1500 DEAs were estimated omitting the input two, then the mean of each scenario was calculated and contrasted against the true efficiency.

III. Results

The idea behind the different correlations structures is basically a matter of DMU's size. If the co–movement between the inputs is positive, a DMU that utilizes a high quantity of x_1 will use a high quantity of x_2, this implies that a DMU that has a high quantity of input one will have a bigger production than a DMU that has a low quantity of input one. On the other hand, if the structure is negative, a DMU that uses a high quantity of x_1 will use a low quantity of x_2, then the DMUs will have a more homogeneous production.

Figure 1 shows the scatter plots of three specific exercises associated with different co–movements. The mean output's variances are 90.0 (positive structure), 50.4 (independent structure) and 10.4 (negative structure).

Figure 1. Scatter plots of different input co–movements structures.



On the other hand, the input correlation structure contains information about the intensity of inputs used. The intensity, measured as (x_1 / x_2), is more homogeneous in positive correlation structures than in negative correlation structures. The standard deviations for some specific exercises are 6.43%, 14.75% and 20.40% for the positive, independent and negative correlation structures, respectively. This implies, in the case of a positive correlation structure, a complementary production process, while a negative correlation structure implies a substitutive production process.

Given that assuming different correlation structures has implications for the DMU's homogeneity, it is better to use the variable return to scale version of DEA, because the DMUs simulated can be in different states of the production process.

Table 1 shows the median for each scenario. As seen, the problem of omitting a relevant variable in the analysis causes inconsistency in the outcomes that are obtained from DEA (the true efficiency is 100%). This wrong specification is more problematic in the case of a negative correlation structure. This means that in the case of a positive correlation, an input incorporates approximately the information contained in the other one, but if a relevant input is omitted, and this is negatively correlated, a relatively efficient DMU can be drastically judged as inefficient.

Table 1. Mean of efficiencies calculated by DEA when a relevant input is omitted: different correlation structures between the inputs.

	CORRELATION STRUCTURE		
	Positive	**Independent**	**Negative**
Mean of Efficiency	91.3%	81.5%	77.3%

IV. Conclusions

The correlation structure of the inputs has implications over the homogeneity of the decision making units; specifically, a positive correlation structure implies a complementary production process and bigger differences in production size. On the other hand, a negative structure means a substitutive process and smaller differences in production. If the correlation between inputs is positive, the set of decision making units is more heterogeneous signifying it is better to use the variable return to scale version of DEA.

Omitting relevant variables in DEA causes an inconsistency in the outcomes. The problem is worse, if the correlation structure of inputs is negative. Given a positive correlation structure between inputs, the efficiencies calculated are closer to true efficiencies, this implies that one variable incorporates information about another one; however, omitting a relevant variable that is negatively correlated with other inputs, causes to judge inefficient a decision making unit that is efficient, which has serious implications; for example, if a regulator uses DEA as a tool to evaluate efficiency, a bad specification can cause huge losses to an efficient decision making unit.

References

Banker, R.(1993). “Maximum Likelihood, Consistency, and Data Envelopment Analysis: A Statistical Foundation”. Management Science 39(10), pags. 1265–1273.

Banker, R. (1989). “An introduction to data envelopment analysis with some of its models and their uses”. Research in government and nonprofit accounting, 5: 125–163.

Banker, R., Charnes, A. and Cooper, W. (1984). “Models for Estimating Technical and Scale Efficiencies in Data Envelopment Analysis”. Management Science 30(9).

Charmes, A., Cooper, W. and Rhodes, E. (1978). “Measuring Efficiency of Decision Making Units”. European Journal of Operations Research, 3, 4, July.

Charmes, A., Cooper, W. and Rhodes, E. (1979). “Short Communication: Measuring Efficiency of Decision Making Units”. 3, 4, July.

Charmes, A., Cooper, W. and Rhodes, E. (1981). “Evaluating Program and Managerial Efficiency: An Application of Data Envelopment Analysis to Program Follow Through”. Management Science, 27, 6, June.

Charmes, A., Cooper, W., Huang, Z. and Sun, D. (1990). “Polyhedral Cone–Ratio DEA Models with an Illustrative Application to Large Commercial Banks”. Journal of Econometrics, 46 pags. 73–91.

Charmes, A., Cooper, W., Thrall, R. (1991). “A Structure for Characterizing and Classifying Efficiency and Inefficiency in Data Envelopment Analysis”. Journal of Productivity Analysis 2, 197–237.

Färe, R. and GrossKopf, S. (1985). “A Nonparametric Cost Approach to Scale Efficiency”. Scandinavian Journal Economics 87: 594–604.

Pedraja, F., Salinas, J. and Smith, P. (1999). “On the Quality of the Data Envelopment Analysis Model”. The Journal of Operational Research Society, 50, 6: 636–644.

Thompson, R., Langemeier, L., Lee, C. y Thrall, R. (1990). “The Roll of Multiple Bounds in Efficiency of Analysis with Application to Kansas Farming”. Journal of Econometrics, 46 pags. 93–108.