Computers & Operations Research 37 (2010) 1087–1090

Contents lists available at ScienceDirect

Computers & Operations Research

journal homepage: www.elsevier.com/locate/caor

Three-stage DEA models for incorporating exogenous inputs

Sarah M. Estelle [a], Andy L. Johnson [b], John Ruggiero [c,*]

[a] Rhodes College, United States
[b] Texas A&M, United States
[c] University of Dayton, Dayton, OH 45469-2251, United States

ARTICLE INFO

Available online 4 October 2009

Keywords:
Data envelopment analysis
Non-discretionary inputs

ABSTRACT

In this paper, we discuss three-stage models that control for exogenous, non-discretionary inputs in data envelopment analysis. In a recent article in this journal, Monte Carlo analysis was employed to compare and contrast alternative DEA models that measure efficiency in the presence of exogenous variables. The methodology for comparison was flawed, calling into question the results presented. We introduce new second-stage models and compare and contrast them with simulated data.

1. Introduction

Data envelopment analysis (DEA) was developed to measure the performance of decision-making units in multiple input, multiple output settings. The seminal papers of Charnes et al. [17] and Banker et al. [1] introduced the constant returns to scale and variable returns to scale models, respectively. DEA, which measures the reciprocal of the distance function introduced by Shephard [18], uses linear programming to identify a linear approximation of the production frontier to allow benchmarking and performance evaluation. The underlying model, however, did not consider non-discretionary inputs, i.e. inputs beyond the control of the producing unit. Banker and Morey [2] introduced a one-stage model to control for these exogenous variables. The Banker and Morey model assumed convexity with respect to the non-discretionary inputs. However, when modeling exogenous variables the assumption of convexity might be invalid as noted by Ruggiero [13]. Ruggiero provided an alternative model that relaxed this constraint. However, this model limits the reference set for identifying benchmark performance based on the exogenous variable. When there are multiple exogenous variables this model suffers from the curse of dimensionality and can no longer discriminate inefficient performance.

To address this issue Ruggiero [14] introduced a three-stage model. In the first stage, the distance function is measured relative to discretionary inputs and outputs. The resulting index captures not only inefficiency but the effect that exogenous variables have on production. In the second-stage, OLS is used to control the influence of exogenous factors. The coefficients obtained from the regression are used to aggregate the multiple exogenous variables to construct an index of environmental influence; in a third-stage, this index is then used in the Ruggiero [13] model. Essentially, the second-stage model decomposes inefficiency and environmental influence and the third-stage measures efficiency maintaining the desirable properties of the DEA measures.

Alternative models have been developed. Ray [12] used a two-stage model to estimate the distance function in the first stage, followed by ordinary least-squares regression in the second stage. Ray uses the error term from the second stage as a measure of efficiency. Muñiz [8] provided an alternative model using a distance function in the first stage. However, the focus in Muñiz's approach is on slacks as opposed to the equiproportional measure. Muñiz argues that first stage slack results from either technical efficiency or from non-controllable factors. More recently, however, Johnson and Ruggiero [6] show that the focus on remaining slack after Farrell efficiency is achieved is misguided. In particular, they show that a benchmark exists in the neighborhood of the Farrell projection that has no additional slack.

Muñiz et al. [9] compared several methods for controlling for exogenous variables by simulating a production process and varying the number of non-discretionary factors. The performance of the methods was compared according to the rank correlation and MAD between estimated and true efficiency. The results indicated that the three-stage model of Ruggiero [14] performed best in nearly all model scenarios and was the only model robust to sample size and the number of non-discretionary variables.

In a recent paper in this journal, Cordero et al. [5] compare and contrast alternative approaches for accounting for non-discretionary variables in DEA. In particular, the performance of one-, two- and three-stage models is analyzed via a Monte Carlo simulation. Their selection of these DEA models is curious. For the two-stage models, the authors discuss using OLS and tobit in a second-stage regression. The authors include a discussion from Simar and Wilson [15] that claims the second-stage regression

[*] Corresponding author.
E-mail address: ruggiero@notes.udayton.edu (J. Ruggiero).

0305-0548/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cor.2009.09.015

parameters are biased. However, more recently, Banker and Natarajan [3] and McDonald [7] prove that OLS provides consistent estimates in the second-stage regression; McDonald [7] also shows that tobit is not consistent with the data generating process.

In this paper, we provide an alternative assessment of the non-discretionary DEA models. In the next section, we constructively review Cordero et al. [5]. In Section 3, we present the models that will be used in our Monte Carlo analysis. Further, new second-stage methods using alternative regression specifications are introduced into Ruggiero [14] three-stage approach. In particular, fractional logit and nonparametric regression are considered as alternatives to OLS. Section 4 presents a Monte Carlo analysis, and the last section concludes.

2. Cordero, Pedraja and Santin (*CPS*, 2009)

The main purpose of *CPS* is to compare alternative methods for treating exogenous variables via Monte Carlo simulation. However, the specification of the data generating process causes concern. The authors argue that previous research designs are flawed because the data generating process (DGP) does not use a flexible functional form. However, the importance of flexible functional forms lies in estimation where the true technology is unknown and not in data generation. Since DEA is nonparametric, it only requires that the DGP be consistent with the axioms of DEA (i.e., convexity, free disposability, and minimum extrapolation) (see Banker et al. [1] for further discussion). The selection of a particular functional form for Monte Carlo analysis is arbitrary and selecting a flexible functional form is no better than selecting any other functional form that is consistent with axioms of DEA.

The following two-discretionary input (x_1, x_2) and one output (y) translog production function is specified in *CPS*:

$$\ln(y) = \beta_0 + \beta_1 \ln(x_1) + \beta_2 \ln(x_2) + 0.5\beta_{11}[\ln(x_1)]^2 + 0.5\beta_{22}[\ln(x_2)]^2 + 0.5\beta_{12} \ln(x_1)\ln(x_2) - u, \tag{1}$$

where $u = -\ln(^{-W} + g(z_1, z_2) + v)$ is a composite term that captures inefficiency W, the effect of non-discretionary factors $g(z_1, z_2)$, and measurement error v. It is possible that u is undefined given the specified data generating process. For example, with v=0 an efficiency value of 0.5 (i.e., W=0.693) causes u to be undefined. The authors have not described what to do in the case of u undefined.

Further, it is not clear why measurement error should be included in u. The authors only consider DEA-based methods that are deterministic. Since *CPS* does not analyze the effect that measurement error has on measured performance, it is not clear why measurement error was included in the data generating process. Furthermore, given the specification of the distribution of v, $v \sim N(0, 0.02)$, it is not clear that the noise is large enough to have any impact on the analysis. In fact, measurement error only accounts for less than 0.1 percent of the variation in $\ln(y)$.

The biggest problem with the DGP is the choice of production function. *CPS* specifies a translog production function, but assigns parameter values that are contrary to the axioms of DEA and, therefore, inappropriate. In particular, by assuming that all parameters are positive, *CPS* violates the convexity requirement of DEA. We show the implications of the assumed functional form by focusing on the scale elasticity. Given (1), the scale elasticity is

$$\varepsilon = \sum_{i=1}^{2} \beta_i + \sum_{i=1}^{2} \sum_{j=1}^{2} \beta_{ij} \ln(x_i). \tag{2}$$

Using the parameters chosen by the authors, we obtain:

$$\varepsilon = 0.6 + 0.2\ln(x_1) + 0.2\ln(x_2). \tag{3}$$

The authors generate $x_i \sim U(1,50)$. Without loss of generality, we will analyze the scale elasticity for the hyperplane defined by x_1=x_2=x. We can rewrite (3) as

$$\varepsilon = 0.6 + 0.4\ln(x), \tag{4}$$

obtaining

$$\frac{d\varepsilon}{dx} = \frac{0.4}{x} > 0. \tag{5}$$

As x and, hence y, increases, scale elasticity increases. We obtain decreasing, constant and increasing returns to scale when x is less than, equal to, or greater than e, respectively. This is in sharp contrast to economic production theory and would be equivalent to a "Law of Increasing Marginal Product". In the context of DEA, the axiom of convexity is violated.

CPS incorporates two exogenous variables z_1 and z_2 that are generated from a uniform distribution on the range from -0.25 to +0.25. The authors give the justification as "(the) effect on observed inefficiency can be both positive or negative." A positive or negative effect is determined by the derivative of output with respect to the non-discretionary variable and not the value. For the DEA approaches such as the one-stage method of Banker and Morey, the models are translation invariant (see [11]). Therefore, the inclusion of negative values for z has no impact on the efficiency estimates of the models. Further for the multi-stage approaches that use regression, the coefficient of the z-variable is robust to the interval selected as long as the interval has the same span. Therefore, z variables generated on any 0.5 span would have similar results.

CPS's discussion of the two-stage method reinforces several misconceptions about the method. In describing the two-stage process proposed by Ray [12], the paper recognized that the second-stage regression parameters can be biased "due to the fact that efficiency scores calculated in (the) first stage,…, depend on all observed inputs and outputs." The comment recognized that if the non-discretionary inputs effect efficiency, this information should be used in the efficiency estimate procedure to improve the efficiency estimates. This is similar to the omitted variable bias in regression. However, the paper fails to recognize that if the correlation between the discretionary inputs and non-discretionary inputs is zero then a two-stage approach will not be biased. Rather, the next paragraph states "this problem can be overcome by using bootstrap methods." The bootstrapping methods correct for sampling bias which has nothing to do with the omitted variable bias discussed in the previous paragraph. Further, the Monte Carlo analysis performed in the paper assumes the correlation between the discretionary inputs and non-discretionary inputs is zero. Thus the discussion of the omitted variable bias in the first stage is unnecessary.

3. DEA models with exogenous inputs

In this section, we describe several methods for handling efficiency in the presence of exogenous variables. Further, we develop new alternative approaches to the measurement of efficiency in the presence of exogenous inputs. We assume that each decision-making unit (DMU) i=1,…,N produces a vector of S outputs y_i=(y_{i1},…,y_{is}) with a vector of M inputs $x_i = (x_{i1}, \ldots, x_{iM})$ given a vector of R non-discretionary inputs $z_i = (z_{i1}, \ldots, z_{iR})$. For this section, we will assume for convenience that $z_2 > z_1$ implies $y_2 > y_1$ holding inputs constant.

The Banker and Morey [2] variable returns to scale model for the measurement of efficiency of DMU "0" is given by

$$\begin{aligned} & BMV_0 = \min\theta \\ \text{s.t.} \quad & \sum_{i=1}^{N}\lambda_i y_{ij} \geq y_{0j,} \quad j=1,\ldots,S, \\ & \sum_{i=1}^{N}\lambda_i x_{ik} \leq \theta x_{0k,} \quad k=1,\ldots,M, \\ & \sum_{i=1}^{N}\lambda_i z_{il} \leq z_{0l,} \quad l=1,\ldots,R, \\ & \sum_{i=1}^{N}\lambda_i = 1, \\ & \lambda_i \geq 0. \end{aligned} \tag{6}$$

This model seeks the maximum equiproportional reduction in all discretionary inputs consistent with observed production and non-discretionary inputs. The difference between this model and the standard DEA model is the removal of the efficiency factor from the right-hand side of the non-discretionary inputs. Banker and Morey also provided the constant returns to scale version of (6):

$$\begin{aligned} & BMC_0 = \min\theta \\ \text{s.t.} \quad & \sum_{i=1}^{N}\lambda_i y_{ij} \geq y_{0j,} \quad j=1,\ldots,S, \\ & \sum_{i=1}^{N}\lambda_i x_{ik} \leq \theta x_{0k,} \quad k=1,\ldots,M, \\ & \sum_{i=1}^{N}\lambda_i z_{il} \leq (z_{0l})\sum_{i=1}^{N}\lambda_i, \quad l=1,\ldots,R, \quad \lambda_i \geq 0. \end{aligned} \tag{7}$$

As pointed out by Ruggiero [13], however, these models assume convexity with respect to the non-discretionary inputs.

Ruggiero (1998) provided a three-stage model to control for non-discretionary inputs when $R>1$. In the first stage, in the analysis of DMU "0", the standard DEA model is applied using only outputs and discretionary factors[1]:

$$\begin{aligned} & F_0 = \min\theta \\ \text{s.t.} \quad & \sum_{i=1}^{N}\lambda_i y_{ij} \geq y_{0j,} \quad j=1,\ldots,S, \\ & \sum_{i=1}^{N}\lambda_i x_{ik} \leq \theta x_{0k,} \quad k=1,\ldots,M, \quad \sum_{i=1}^{N}\lambda_i = 1, \quad \lambda_i \geq 0. \end{aligned} \tag{8}$$

The resulting index is composed of inefficiency and the effect that non-discretionary variables have on the production process. In the second stage, the following regression is applied:

$$F_i = \alpha + \boldsymbol{\beta}' z_i + \varepsilon_{i.} \tag{9}$$

Ray [12] applied OLS to estimate (9); McCarty and Yaisawarng [19] recommended tobit given that $0<F_i\leq 1$. Banker and Natarajan [3] and McDonald [7] both prove that OLS is a consistent estimator. Here the functional form is assumed to be linear which is common in the literature.

In this paper, we extend the methodology by considering two alternative regression approaches in the second stage. First, we provide a nonparametric regression, making the three-stage approach entirely nonparametric. We consider the following regression equation:

$$F_i = g(z_i) + \varepsilon_i. \tag{10}$$

The regression function $g(z)$ can be locally approximated by fitting a function to the data points within a chosen neighborhood, a method pioneered by Cleveland [4]. It is common to identify neighborhoods by placing each data point in the data set at the center of its own neighborhood. Weighted least squares is used to fit linear or quadratic functions within a specified neighborhood of the predictors, defined by the smoothing parameter (i.e., the radius of the neighborhood). For our model, we chose linear functions.[2]

We also consider a fractional logit, a parametric alternative to OLS and the nonparametric regression the second stage. Since the dependent variable in the second stage is bounded by 0 and 1, the effect of a given change in an independent variable z on F must vary throughout the range of z. Thus, simple ordinary least squares will predict infeasible F for some combination of independent variables and, therefore, can be inappropriate. Since the frontier in stage one includes some DMUs by construction, F will equal to 1 for some observations. Thus, a straightforward log-odds transformation—which is undefined for values equal to 1—is not a solution to dealing with our proportion data. A better empirical model when the dependent variable is proportional with some observed efficiency value equal to 1 and/or 0 is fractional logit (or, flogit) as suggested by Wooldridge [16].[3]

Nonparametric regression, fractional logit, or ordinary least squares can be used in the second stage is to parse out the effect that non-discretionary inputs have on production. Determining which method works best is an empirical question that will be investigated in the Monte Carlo simulation, presented in this paper. Ruggiero [14] showed that an overall index of environmental harshness can be obtained from $z_i=\hat{F}_i$, the predicted first-stage index. This index is then used as a control variable in a third-stage model[4]:

$$\begin{aligned} & TS_0 = \min\theta \\ \text{s.t.} \quad & \sum_{i=1}^{N}\lambda_i y_{ij} \geq y_{0j,} \quad j=1,\ldots,S, \\ & \sum_{i=1}^{N}\lambda_i x_{ik} \leq \theta x_{0k,} \quad k=1,\ldots,M, \\ & \sum_{i=1}^{N}\lambda_i = 1, \quad \lambda_i \geq 0, \quad \lambda_i = 0 \text{ if } z_i > z_0. \end{aligned} \tag{11}$$

In the evaluation of DMU "0" any DMU with a more favorable environment, represented by a higher index value, is excluded from the solution space.

4. Monte Carlo analysis

Assume that the production technology can be represented by the transformation of two-discretionary inputs (x_1 and x_2) into

[1] We present the variable returns to scale model. For our Monte Carlo analysis, we employed the constant returns to scale version by removing the convexity constraint to be consistent with the underlying technology.

[2] For the applications to simulated data in this paper, we used Proc Loess in SAS. The smoothing parameter is chosen optimally in SAS to satisfy information criteria.

[3] For empirical applications of flogit, it is important to account for potential heteroskedatiscity. STATA software can be used to obtain flogit parameter estimates, robust standard errors, and marginal effects, all of which are useful in interpreting empirical results. However, since we are conducting Monte Carlo simulations, we are only interested in the predicted values of F that follow from the estimated flogit coefficients. For more about empirical applications of flogit, see Papke and Wooldridge [10].

[4] For the Monte Carlo analysis, the constant returns to scale version of this model was employed to be consistent with the data generating process.

Table 1
Monte Carlo results.

Method	Correlation	Rank correlation	MAD
One-stage model			
BMC	0.381 (0.059)	0.408 (0.065)	0.283 (0.035)
Three-stage models			
OLS	0.877 (0.017)	0.821 (0.023)	0.074 (0.003)
Nonparametric	0.875 (0.022)	0.818 (0.025)	0.073 (0.003)
Fractional logit	0.877 (0.017)	0.821 (0.023)	0.073 (0.003)

Reported results are averages (standard deviations) from 100 replications. The three-stage models are based on using linear programming model (8) in the first-stage, regression equation (9) in the second-stage and linear programming model (11) in the third-stage.

output (y) given two non-discretionary inputs (z_1 and z_2) according to the production function

$$y_i = z_{i1}^2 z_{i2}^{-3} x_{i1}^{0.4} x_{i2}^{0.6}, \qquad (12)$$

where constant returns to scale exist.[5] Simulated inputs were drawn for 500 DMUs from the following distributions:

$$\begin{aligned} \ln x_k &\sim N(0,1), \quad k=1,2 \\ \ln z_l &\sim N(0,0.1), \quad l=1,2. \end{aligned} \qquad (13)$$

An inefficiency component $|u|$ was generated with $u \sim N(0, 0.2)$; observed output was generated as $y_i^o = e^{-|u|} y_i$, where $e^{-|u|}$ is a measure of technical efficiency. Finally, 100 replications were performed.

Efficiency was estimated using four models. For the one-stage model, we applied the constant returns to scale Banker and Morey (7) model. In the data generating process, non-discretionary input z_2 has a negative exponent indicating an adverse effect on production; to be consistent with the Banker and Morey model, this variable was inverted so that higher levels represented a better environment. For the three-stage models, constant returns to scale were assumed for the first-stage estimates (8). Three second-stage regression models were considered: OLS, nonparametric regression and flogit. After obtaining the resulting index, the third-stage DEA model (11) was applied.

The results of the analysis are reported in Table 1. Performance of the various models was evaluated using three criteria: correlation, rank correlation and mean absolute deviation (MAD) between true and estimated efficiency. The reported results are the average results from 100 replications. Higher correlations and lower MADs are desirable.

The first conclusion to be drawn from our analysis is the poor performance of the one-stage models. The maintained assumption of convexity with respect to the non-discretionary inputs contributes to the poor estimates. Note these results are consistent with the results found in [14]. Notably, the correlations and rank correlations, all below 0.41, are the lowest of all the models considered. The average MADs are above 0.2 for the Banker and Morey model.

Secondly, the three-stage models appear to work relatively well. In particular, the average correlation (rank correlation) was above 0.87 (0.81) for all models. In addition, the MADs are all relatively low, achieving averages approximately 1/3 of those achieved by the one-stage models. While the one-stage model could not decompose the environment and efficiency effects, the three-stage models perform remarkably well. Furthermore, the three models that used different regression approaches in the second-stage produced very similar results. It appears that the second stage is robust with respect to the selection of OLS, flogit or nonparametric regression.

[5] Returns to scale are defined with respect to discretionary outputs only.

5. Conclusions

In this paper, alternative models for analyzing efficiency in the presence of non-discretionary inputs have been analyzed. Two new three-stage models with different regression procedures applied in the second stage are introduced. In particular, we developed flogit and nonparametric regression as two competing regression methods instead of OLS for the second stage of the three-stage models. The Monte Carlo results indicated that the three-stage models are robust with respect to regression type. Future research to demonstrate under what circumstances each regression approach is advantageous would be valuable.

References

[1] Banker RD, Charnes A, Cooper WW. Some models for estimating technical and scale inefficiencies in data envelopment analysis. Management Science 1984;30(9):1078–92.

[2] Banker R, Morey R. Efficiency analysis for exogenously fixed inputs and outputs. Operations Research 1986;34(4):513–21.

[3] Banker R, Natarajan R. Evaluating contextual variables affecting productivity using data envelopment analysis. Operations Research 2008;56(1):48–58.

[4] Cleveland WS. Robust locally-weighted regression and smoothing scatter-plots. Journal of the American Statistical Association 1979;74:829–36.

[5] Cordero JM, Pedraja F, Santín D. Alternative approaches to include exogenous variables in DEA measures: a comparison using Monte Carlo. Computers and Operations Research 2009;36:2699–706.

[6] Johnson A, Ruggiero J. E-substitutability, slacks, and data envelopment analysis. European Journal of Operational Research. 2009, submitted for publication.

[7] McDonald J. Using least squares and tobit in second stage DEA analyses. European Journal of Operational Research 2009;197:792–8.

[8] Muñiz M. Separating managerial inefficiency and external conditions in data envelopment analysis. European Journal of Operational Research 2002;143(3):625–43.

[9] Muñiz M, Paradi J, Ruggiero J, Yang Z. Evaluating alternative DEA models used to control for non-discretionary inputs. Computers and Operations Research 2006;33:1173–83.

[10] Papke L, Wooldridge J. Econometric methods for fractional response variables with an application to 401(K) plan participation rates. Journal of Applied Econometrics 1996;11(6):619–32.

[11] Pastor JT. Translation invariance in data envelopment analysis: a generalization. Annals of Operations Research 1996;66:93–102.

[12] Ray SC. Resource-use efficiency in public schools: a study of Connecticut data. Management Science 1991;37(12):1620–8.

[13] Ruggiero J. On the measurement of technical efficiency in the public sector. European Journal of Operational Research 1996;90:553–65.

[14] Ruggiero J. Non-discretionary inputs in data envelopment analysis. European Journal of Operational Research 1998;111:461–9.

[15] Simar L, Wilson PW. Estimation and inference in two-stage, semi-parametric models of production processes. Journal of Econometrics 2007;136(1):31–64.

[16] Wooldridge J. Econometric analysis of cross section and panel data, 3rd ed. MIT Press; 2002 661–3.

[17] Charnes A, Cooper WW, Rhodes E. Measuring the efficiency of decision making units. European Journal of Operational Research 1978;2:429–44.

[18] Shephard R. Theory of cost and production functions. Princeton, NJ:Princeton University Press, 1970.

[19] McCarty T, Yaisawarng S. Technical efficiency in New Jersey school districts. In: Fried Lovell, Schmidt (editors). The measurement of productive efficiency. New York: Oxford University Press; 1993.