



**EVALUATING CONTEXTUAL VARIABLES AFFECTING PRODUCTIVITY
USING DATA ENVELOPMENT ANALYSIS**

Journal:	<i>Operations Research</i>
Manuscript ID:	OPRE-2005-12-146
Manuscript Type:	Article
Date Submitted by the Author:	21-Dec-2005
Complete List of Authors:	Banker, Rajiv; Temple University, Fox School of Business Natarajan, Ram; UT Dallas, School of Management
Keywords:	Effectiveness/performance < Organizational studies, Productivity < Organizational studies

powered by ScholarOne
Manuscript Central™

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

**EVALUATING CONTEXTUAL VARIABLES AFFECTING
PRODUCTIVITY USING DATA ENVELOPMENT ANALYSIS**

Rajiv D. Banker

Fox School of Business and Management, Temple University,
Philadelphia, Pennsylvania 19122
Banker@temple.edu

Ram Natarajan

School of Management, The University of Texas at Dallas,
Richardson, Texas 75083
Nataraj@utdallas.edu

Last Revised: December 19, 2005

EVALUATING CONTEXTUAL VARIABLES AFFECTING PRODUCTIVITY USING DATA ENVELOPMENT ANALYSIS

Abstract

A DEA-based stochastic estimation framework is presented to evaluate contextual variables affecting productivity. Conditions are identified under which a two-stage procedure consisting of DEA followed by regression analysis yields consistent estimators of the impact of contextual variables. Conditions are also identified under which DEA in the first stage followed by maximum likelihood estimation in the second stage yields consistent estimators of the impact of contextual variables. Monte Carlo simulations are carried out to compare the performance of our two-stage approach with one-stage and two-stage parametric approaches. Simulation results suggest that DEA-based procedures perform as well as the best parametric method in the estimation of the impact of contextual variables on productivity. Simulation results also indicate that DEA-based procedures perform better than parametric methods in the estimation of individual decision making unit (DMU) productivity.

(Key words: Data envelopment analysis, Stochastic frontier analysis, Data generating process, Contextual variables, Two-stage estimation, Monte Carlo simulation)

1. Introduction

More than twenty-five years since its inception, data envelopment analysis (DEA) is used extensively in many management disciplines to analyze the efficiency of organizations. Several years ago, Schmidt (1985) classified DEA as a non-statistical approach and posed the following challenge:

“I am very skeptical of non-statistical measurement exercises, certainly as they are now carried out and perhaps in any way in which they could be carried out....I see no virtue whatever in a non-statistical approach to data.”

Responding to this call, Banker (1993) provided a formal statistical foundation for DEA by identifying conditions under which DEA estimators are statistically consistent and maximize likelihood. Continuing in a similar spirit, we provide a formal statistical basis for the many studies that have used the DEA efficiency score in a second stage analysis to assess its cross-sectional association with contextual variables or socio-economic factors.

Analysis of factors contributing to productivity differences has been an important area of research in DEA. Ray (1991), for instance, regresses data envelopment analysis (DEA) scores on a variety of socio-economic factors to identify key performance drivers in school districts. The two-stage approach of first calculating productivity scores and then seeking to correlate these scores with various explanatory variables has been in popular use (Forsund 1999). Observing that explanations of productivity differences are still dominated by ad hoc speculations, Forsund (1999) emphasizes the need for research into its theoretical underpinnings. Grosskopf (1996) finds the lack of articulation of a data generating process (DGP) consistent with the two-stage analysis to be particularly troubling.

In a typical two-stage study based on DEA, the relative productivity of each organization is evaluated in the first stage based on data on input consumption and output production. The

1
2
3 productivity score is then regressed on potential contextual factors in the second stage to identify
4
5 the factors whose impact on productivity is statistically significant. If, additionally, information
6
7 on managerial performance is needed, the second stage analysis can be extended to estimate
8
9 individual decision making unit (DMU) productivity after filtering the component associated
10
11 with the contextual factors. Alternative second stage methods have included the use of the
12
13 relative productivity score or its logarithmic transform as the dependent variable in an ordinary
14
15 least squares (OLS) regression.
16
17
18
19

20 Earlier studies have not checked whether such a two-stage approach is statistically valid
21
22 for identifying significant contextual variables. Most of the studies in parametric productivity
23
24 analysis that have compared the performance of one-stage and two-stage approaches in analyzing
25
26 the impact of contextual variables on productivity assume that the parametric functional form
27
28 characterizing the production function is known. For example, Wang and Schmidt (2002)
29
30 consider the estimation of a stochastic frontier model under the maintained assumption that the
31
32 underlying parametric functional form is known and argue that two-stage procedures produce
33
34 biased estimates of the effect of environmental factors on efficiency. Monte Carlo evidence in
35
36 their paper supports their argument.
37
38
39
40

41 Often, it may be the case that no a priori knowledge exists about the form of the
42
43 production correspondence except that it is monotone increasing and concave in its inputs.
44
45 Assumption of a specific parametric form may, therefore, influence efficiency estimation as well
46
47 as introduce estimation errors in the one-stage process. In such situations, it may be necessary to
48
49 use a nonparametric method such as DEA to estimate the production correspondence between
50
51 the inputs and the outputs in the first-stage, and follow this procedure with OLS regression or
52
53 maximum likelihood estimation methods in the second stage. DEA estimators exhibit the
54
55
56
57
58
59
60

1
2
3 desirable asymptotic property of consistency (Banker 1993) and the asymptotic distribution of
4 the DEA estimators is identical to the true distribution of the efficiency. Therefore, it is
5 reasonable to expect that the problems associated with using efficiency estimators from a first
6 stage analysis of outputs and inputs will be less acute if DEA rather than a parametric approach
7 is used in the first stage. This is the primary issue we investigate in this study.
8
9

10
11
12
13
14
15 In contrast to many of the prior studies that have exclusively focused on assessing the
16 impact of contextual variables on productivity, we also examine the estimation of DMU-specific
17 productivity attributable to factors within managerial control. In many practical applications,
18 this information is useful to evaluate managerial performance and to provide guidance on the
19 extent to which DMU performance can be improved through better management practices.
20
21
22
23
24
25

26
27 In the analysis that follows, we first provide theoretical justification for the use of a two-
28 stage method that uses DEA in the first stage. We also tabulate extensive Monte Carlo evidence
29 that helps understand the performance of the two-stage method based on DEA in comparison to
30 the various one-stage and two-stage parametric methods advocated in the prior literature.
31 Simulation results suggest that our two-stage DEA-based procedures perform significantly better
32 than one-stage parametric methods that rely on commonly used parametric functional forms such
33 as translog and Cobb-Douglas to specify the production correspondence, which in turn
34 outperform the two-stage parametric methods to evaluate the impact of contextual variables on
35 productivity. Simulation results also indicate that DEA-based methods perform significantly
36 better than one-stage and two-stage parametric methods in the estimation of individual DMU
37 productivity.
38
39
40
41
42
43
44
45
46
47
48
49
50
51

52
53 We begin by specifying output as a general function of inputs and an error term. We
54 model the error term as consisting of three distinct components, a linear function of contextual
55
56
57
58
59
60

1
2
3 variables, a one-sided inefficiency term and a two-sided random noise term bounded above.
4
5 Except for the additional component involving the contextual variables, our specification of the
6
7 error term is analogous to composed error term formulations in parametric stochastic frontier
8
9 models. We specify the production function relating output and inputs as monotone increasing
10
11 and concave. We develop two-stage estimation methods by adapting the DEA+ method
12
13 introduced by Gstach (1998). Consistent estimators of the impact of the contextual variables are
14
15 obtained with DEA in the first stage and OLS or MLE in the second stage.
16
17
18
19

20 We perform simulations in a single output, single input and a single contextual variable
21
22 framework. We use a cubic polynomial specification, monotone and concave within the range of
23
24 the input, to characterize the production function linking the input to the output. We then
25
26 combine realizations of the input, contextual variable, one-sided efficiency and two-sided noise
27
28 from known distributions with the specified production function to generate values of the output.
29
30 We compare a total of twelve methods to evaluate the impact of contextual variable on
31
32 productivity and estimate individual DMU productivity. The methods differ in terms of whether
33
34 a one-stage or two-stage approach is used, whether a specific functional form is used to
35
36 characterize the production function and whether OLS or ML estimation is used. Metrics such as
37
38 mean absolute deviation and mean squared error evaluate the performance of the various
39
40 methods in terms of their ability to determine the influence of the contextual variable on
41
42 productivity.
43
44
45
46
47

48 The paper proceeds as follows. The next section describes our basic model and data
49
50 generating process. Section 3 contains the estimation methods and statistical tests for the two-
51
52 stage method. Section 4 presents the simulation results and section 5 summarizes and concludes.
53
54
55
56
57
58
59
60

2. Basic Framework

Consider observations on $j = 1, \dots, N$ decision making units (DMUs), each observation comprising a vector of outputs $Y_j \equiv (y_{1j}, \dots, y_{Rj})$, a vector of inputs $X_j \equiv (x_{1j}, \dots, x_{Ij})$, and a vector of contextual variables $Z_j \equiv (z_{1j}, \dots, z_{Sj})$ that may influence the overall productivity in transforming the inputs into the outputs, with the non-negative vectors Y_j , X_j and Z_j being strictly positive in at least one dimension. Thus, for instance, in Farrell's (1957) original setting for efficiency evaluation, the outputs Y may represent a farm's production measured in tons of grain, the inputs X may be labor, capital and materials, and the contextual variables Z that influence productivity may be factors such as farm ownership and management methods.¹

We describe our basic model for the case of a single output, y , to maintain direct contact with parametric stochastic frontier models. The extension to the multiple outputs case is straightforward.² The model we specify includes the true production function $\phi(X)$ and an error term ε . The production function $\phi(\cdot)$ is monotone increasing and concave in X , and relates the vector X to a single output y as specified by the equation

$$y = \phi(X) * e^{\varepsilon^*} \quad (1)$$

The random variable representing the error ε^* is itself generated by the process

$$\varepsilon^* = v - u - \sum_{s=1}^S \beta_s z_s \quad (2)$$

where v represents random noise and has a two-sided distribution, u represents technical inefficiency and has a one-sided distribution and the contextual variables z_s are all positive. The

¹ Our definition of contextual variables includes those variables that may be exogenously fixed as well as others that may be under the control of the DMU managers.

² Our extension to the multiple output case involves an additional vector of random variables specifying the proportion of each output. The data generating process then determines the output vector Y_j as in the single output case on the ray defined by the vector of random variables specifying the output mix.

contextual variables are measured such that the weights β_s , $s = 1, \dots, S$ are all non-negative i.e., the higher the value of the contextual variables the higher is the inefficiency of the DMU. Except for the additional component involving the contextual variables, our specification of the error term is analogous to composed error formulations in parametric stochastic frontier models. For the purposes of exposition, we also define the error attributable to only noise and technical inefficiency as $\varepsilon = v - u$.

We impose the following structure on the probability density functions generating the various variables:

$$f_{x_i}(x_i) = 0 \text{ for all } x_i < 0 \quad (3a)$$

$$f_{z_s}(z_s) = 0 \text{ for all } z_s < 0 \quad (3b)$$

$$f_u(u) = 0 \text{ for all } u < 0 \quad (3c)$$

$$f_v(v) = 0 \text{ for all } |v| > V^M \quad (3d)$$

Further, the probability density functions $f_{x_i}(x_i)$, $f_{z_s}(z_s)$, $f_u(u)$ and $f_v(v)$ are all independent of each other.³ Each stochastic variable has finite variance and the mean of the noise variable, $E(v)$, is zero. It is straightforward to verify that the p.d.f. of the composed error, ε , is given by

$$f_\varepsilon(\varepsilon) = \int_{\varepsilon}^{V^M} f_v(v) f_u(v - \varepsilon) dv = \int_0^{V^M - \varepsilon} f_u(u) f_v(u + \varepsilon) du \quad (4)$$

2.1 Estimation of impact of contextual variables on productivity with parametric production function specification

If $\phi(X)$ can be specified as $\phi(X; \gamma)$, where γ is a parameter vector, the relationship in (1) can be transformed to

³ We only need to assume that the input variable vector X , the contextual variable vector Z , the inefficiency u and the noise v are independently distributed. No restrictions are imposed on the joint distribution of the component random variables within the input vector X or the contextual variable vector Z .

$$\ln y = \ln \phi(X; \gamma) - \sum_{s=1}^S \beta_s z_s + \varepsilon \quad (5)$$

For instance, if $\phi(\cdot)$ is Cobb-Douglas, then

$$\ln y = \gamma_0 + \sum_{i=1}^I \gamma_i \ln X_i - \sum_{i=1}^S \beta_i z_i + \varepsilon \quad (6a)$$

or, if $\phi(\cdot)$ is translog, then

$$\ln y = \gamma_0 + \sum_{i=1}^I \gamma_i \ln X_i + \sum_{i=1}^I \gamma_{ii} (\ln X_i)^2 + \sum_{\substack{i,j=1 \\ i < j}}^I \gamma_{ij} \ln X_i \ln X_j - \sum_{i=1}^S \beta_i z_i + \varepsilon \quad (6b)$$

For the data generating processes described in equations 3(a)-3(d) together with (6a) or (6b), two types of estimation procedures are possible. OLS can be used since ε has a finite variance although it may not have a zero mean (Schmidt 1976). The OLS estimators of β_i are consistent and the corresponding t-statistics can be used to evaluate whether a particular contextual variable impacts productivity.

Alternatively, MLE can be used if a specific parametric form is assumed for the p.d.f. of ε . For instance, if v is assumed to be distributed as Normal and u as either half-Normal or exponential as in Aigner et al (1977) or Meeusen and van den Broeck (1997), then $V^M = \infty$, the domain of the composed error ε is $(-\infty, \infty)$, the ML estimators of β_i are consistent, asymptotically normally distributed and efficient and the corresponding t-statistics can be used to evaluate whether a particular contextual variable impacts productivity. The relative contribution of the noise v and the inefficiency u can also be evaluated. If estimators of individual inefficiencies are desired, conditional estimators of u , based on observed values of ε , can be estimated as in Jondrow et al. (1982).

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Some researchers (Pitt and Lee 1981, Kalirajan 1989) have employed a two-stage parametric approach where the relationship $\ln y = \ln \phi(X; \gamma) + \varepsilon^*$ is estimated in the first stage, and the relationship involving the estimated residual with the contextual variables, $\hat{\varepsilon}^* = - \sum_{i=1}^S \beta_i z_i + \varepsilon$ is estimated in the second stage. However, such two-stage parametric approaches suffer from biased estimation of β in the second stage. Wang and Schmidt (2002) explain that the problem is due to the estimator of technical efficiency u , derived from the first stage residual, having a “shrinkage” towards the mean and as a result showing less dependence on the contextual variables \mathbf{z} than it should. Therefore, the estimated weights on the contextual variables end up being smaller than the actual weights. The evidence from the simulations in Wang and Schmidt (2002) confirms that this bias is substantial. Therefore, for the parametric approach, a single stage parametric procedure that jointly estimates inefficiency and the impact of contextual variables is the appropriate approach for productivity analysis.

3. Estimation of impact of contextual variables on productivity for general monotone increasing, concave production functions

Frequently it is the case that no guidance on a parametric specification is available on the production function $\phi(\cdot)$ and a priori all that may be known about $\phi(\cdot)$ is that it is monotone increasing and concave in X . The issue is whether a two-stage estimation procedure involving DEA in the first stage followed by OLS or ML estimation in the second stage suffers from the same problem that affects the two-stage parametric procedure.

Specifically, we propose the use of DEA in the first stage to estimate individual DMU productivity combined with a parametric approach in the second stage to evaluate the contextual

variables affecting productivity. The first stage involves a DEA-based estimation procedure similar in spirit to the DEA+ method introduced by Gstach (1998). We define:

$$\ln \tilde{\phi}(X) = \ln \phi(\cdot) + V^M \text{ and} \quad (7a)$$

$$\begin{aligned} \ln \tilde{\theta} &= (\varepsilon - V^M) - \sum_{i=1}^S \beta_i z_i \\ &= (v - V^M) - u - \sum_{i=1}^S \beta_i z_i \leq 0 \end{aligned} \quad (7b)$$

Since $\tilde{\phi}(X)$ is derived from $\phi(\cdot)$ by multiplication with a positive constant, $\tilde{\phi}(X)$ is also monotone increasing and concave. Substituting (7a) and (7b) into (1) yields

$$\ln y = \ln \tilde{\phi}(X) + \ln \tilde{\theta} \quad (8)$$

Since $\tilde{\theta} \leq 1$, this resembles the usual DEA model (e.g., Banker 1993). Therefore, we refer to the deviation of observed output y from the derived frontier $\tilde{\phi}(X)$ as productivity $\tilde{\theta}$. The logarithm of the DEA productivity estimator, $\ln \hat{\theta}$, obtained by performing DEA on the input-output observations (X_j, y_j) , $j = 1, \dots, N$, is a consistent estimator of $\ln \tilde{\theta}$ (Banker 1993). As observed earlier, these concepts extend directly to the multi-output case. See, also, Banker, Janakiraman and Natarajan (2002) for statistically consistent estimation of general monotone and concave or convex functional relationships in composed error situations.

Comparing (1), (7b) and (8), we note that the productivity from the first stage, $\tilde{\theta} \leq 1$, can be related to the contextual variables by the relation:

$$\ln \tilde{\theta} = - \sum_{i=1}^S \beta_i z_i - \tilde{\varepsilon} \quad (9)$$

where $\tilde{\varepsilon} = V^M - \varepsilon \geq 0$. A comparison of (9) with traditional parametric specification of production frontiers (Aigner and Chu 1968) indicates that the two specifications are very similar

although the independent and dependent variables in (9) are not outputs or inputs. OLS and MLE based methods have been suggested as being appropriate for parametric production frontier estimation. In a similar spirit, we next propose OLS and MLE based estimation procedures that provide consistent estimators of the parameter vector β in (9).

3.1 Using OLS in the second stage

Define $\beta_0 = E(\varepsilon) - v^M$ and $\delta = \varepsilon - E(\varepsilon)$. Now, (9) can be rewritten as

$$\ln \tilde{\theta} = \beta_0 - \sum_{i=1}^s \beta_i z_i + \delta \quad (10)$$

Evidently, the error term δ in (10) has a zero mean and a finite variance. Therefore, if $\tilde{\theta}$ is known, OLS estimation of (10) yields consistent estimators of all components of β and the OLS estimator of the intercept provides a consistent estimator of $E(\varepsilon) - v^M$ (Schmidt 1976).

Since the true productivity is not known in the second stage estimation, we replace $\ln \tilde{\theta}$, the dependent variable in (10), with the corresponding DEA estimator, $\ln \hat{\theta}$. It is intuitive that as far as evaluating the impact of the individual contextual variables is concerned, the use of $\ln \hat{\theta}$ instead of the true $\ln \tilde{\theta}$ preserves consistency.

Proposition 1: If $\mathbf{Q} = p \lim(Z'Z/n)$ is a positive definite matrix, then the OLS estimator of $\tilde{\beta}$ in

$$\ln \hat{\theta} = \tilde{\beta}_0 - Z\tilde{\beta} + \tilde{\delta} \quad (11)$$

yields a consistent estimator of the corresponding components of the parameter vector β .

Proof: See the Appendix.

Therefore, if the objective of the analysis is to evaluate the impact of the contextual variables, then DEA followed by OLS is valid under the maintained assumptions of our data generating process.

3.2 Using MLE in the second stage

There are two aspects that need to be considered in determining whether ML estimation of (10) will yield consistent estimators of the parameters in the second stage. The first issue is related to the fact that standard proofs for consistency of ML estimators are not applicable to ML frontier estimation with bounded error distribution when the support of the distribution itself is endogenously determined. Greene (1980) discusses this issue in detail and provides sufficient conditions on the density of the composed error term under which ML estimation provides consistent estimators. The second issue is related to the fact that the second stage estimation uses the DEA estimator $\hat{\theta}$ in place of the true value $\tilde{\theta}$ and this substitution may impact the properties of the estimators of β .

To address the first issue we propose elementary density functions for u and v that satisfy Greene's conditions. Greene (1980) identifies the following two sufficient conditions on the density of the composed error term ε to ensure that the ML estimator is consistent and its asymptotic distribution is Normal:

$$(i) \quad f_{\varepsilon}(\boldsymbol{\varepsilon}) = \int_{\boldsymbol{\varepsilon}}^{V^M} f_v(v) f_u(v - \boldsymbol{\varepsilon}) dv = 0 \quad \text{at } \boldsymbol{\varepsilon} = V^M$$

$$(ii) \quad \left. \frac{\partial f_{\varepsilon}(\boldsymbol{\varepsilon})}{\partial \boldsymbol{\varepsilon}} \right|_{\boldsymbol{\varepsilon}=V^M} = -f_v(\boldsymbol{\varepsilon}) f_u(0) \Big|_{\boldsymbol{\varepsilon}=V^M} = 0$$

The first condition is always satisfied at $\boldsymbol{\varepsilon} = V^M$. The second condition implies that $f_v(V^M) = 0$ [e.g. Beta] or $f_u(0) = 0$ [e.g. Gamma or Lognormal].

Suppose u is distributed as Gamma $(2, \lambda)$ and v as $N(0, \sigma_v^2)$ truncated above and below, respectively, at V^M and $-V^M$. Let $\boldsymbol{\varepsilon}_1 = \left(\frac{\boldsymbol{\varepsilon}}{\sigma_v} + \frac{\sigma_v}{\lambda} \right)$, $\boldsymbol{\varepsilon}_2 = \left(\frac{V^M}{\sigma_v} + \frac{\sigma_v}{\lambda} \right)$, and $f^*(\cdot)$ and $F^*(\cdot)$ be the standard normal density and distribution functions, respectively. The p.d.f. of $\boldsymbol{\varepsilon} = v - u$ is:

$$f(\boldsymbol{\varepsilon}) = \frac{\sigma_v e^{\frac{\sigma_v^2 + \boldsymbol{\varepsilon}}{2\lambda^2 + \lambda}}}{\lambda^2 \left\{ F^* \left(\frac{V^M}{\sigma_v} \right) - F^* \left(\frac{-V^M}{\sigma_v} \right) \right\}} \left[\left\{ f^*(\boldsymbol{\varepsilon}_1) - f^*(\boldsymbol{\varepsilon}_2) \right\} + \boldsymbol{\varepsilon}_1 \left\{ F^*(\boldsymbol{\varepsilon}_1) - F^*(\boldsymbol{\varepsilon}_2) \right\} \right] \quad (12)$$

Since (9) can also be expressed as

$$\boldsymbol{\varepsilon} = \sum_{i=1}^S \boldsymbol{\beta}_i z_i + V^M + \ln \tilde{\boldsymbol{\theta}} \quad (13)$$

the log-likelihood function can be formed as $\sum_{j=1}^N \ln f \left(\sum_{i=1}^S \boldsymbol{\beta}_i z_{ij} + V^M + \ln \tilde{\boldsymbol{\theta}}_j \right)$ using (13).

Maximizing this log-likelihood function with respect to $\boldsymbol{\beta}$, λ , σ_v and V^M yields consistent estimators of the unknown parameters. As shown in the following proposition, the issue regarding the use of the DEA estimator of the productivity measure $\tilde{\boldsymbol{\theta}}$ rather than the true value of $\boldsymbol{\theta}$ itself is addressed by exploiting Banker's (1993) result on the consistency of the DEA estimator.

Proposition 2: The ML estimator of $\tilde{\boldsymbol{\beta}}$ in

$$\hat{\boldsymbol{\varepsilon}} = Z\tilde{\boldsymbol{\beta}} + V^M + \ln \hat{\boldsymbol{\theta}} \quad (14)$$

yields a consistent estimator of the parameter vector $\boldsymbol{\beta}$.

Proof: See the Appendix.

3.3 Estimating individual inefficiencies

So far we have described how the impact of the contextual variables can be assessed through OLS or MLE in the second stage. In some instances, estimates of the inefficiency variable u are needed to evaluate the performance of individual DMUs. Jondrow et al. (1982) describe the estimation of inefficiency conditional on the observed value of the composed error term for stochastic frontier models discussed in Aigner et al. (1977). We proceed along similar lines to estimate individual inefficiency conditional on the value of the composed error term.

When u is distributed as Gamma $(2, \lambda)$ and v as $N(0, \sigma_v^2)$ truncated truncated above and below, respectively, at V^M and $-V^M$, the conditional distribution of the inefficiency u given ε is :

$$f(u|\varepsilon) = \frac{u e^{-\frac{(u+\varepsilon)}{2\sigma_v^2}} \frac{1}{2\lambda}}{\sqrt{2\pi}\sigma_v^2 \left[\{f^*(\varepsilon_1) - f^*(\varepsilon_2)\} + \varepsilon_1 \{F^*(\varepsilon_1) - F^*(\varepsilon_2)\} \right]} \quad \text{with } 0 \leq u \leq V^M - \varepsilon \quad (15)$$

The conditional mean $E(u|\varepsilon)$ is derived from (15) as $E(t) \left[1 + \frac{\text{Var}(t)}{(E(t))^2} \right]$, where t is distributed as

$N\left(-\left(\varepsilon + \frac{\sigma_v^2}{\lambda}\right), \sigma_v^2\right)$ truncated between 0 and $V^M - \varepsilon$. $E(u|\varepsilon)$ is a consistent estimator of u given

ε . Also, the conditional mode $M(u|\varepsilon)$ is a MLE as in Jondrow et al. (1982). The conditional mode is given by the following:

- If $\varepsilon \geq V^M - \frac{1}{\frac{\sigma_v^2}{\lambda} + \frac{1}{V^M}}$ then $M(u|\varepsilon) = V^M - \varepsilon$ (16a)

- Else, $2M(u|\varepsilon) = -\left(\varepsilon + \frac{\sigma_v^2}{\lambda}\right) + \sqrt{\left(\varepsilon + \frac{\sigma_v^2}{\lambda}\right)^2 + 4\sigma_v^2}$ (16b)

Evidently, both $E(u|\epsilon)$ and $M(u|\epsilon)$ are monotone decreasing functions of ϵ . Therefore, if the objective is only to rank order observations in terms of their estimated inefficiency, the task can be accomplished directly by rank ordering on the basis of $\hat{\epsilon}$ itself.

4. Simulations

In this section we conduct simulations to evaluate the performance of one-stage and two-stage estimation methods in assessing the impact of contextual variables on productivity. Based on prior research, we expect to find that one-stage parametric procedures outperform two-stage parametric methods. We also expect to find that two stage procedures that use DEA in the first stage outperform two-stage parametric methods. Simulation evidence will help us assess the performance of the two-stage DEA-based procedures relative to the one-stage parametric estimation methods.

4.1 Design of the Experiment

We represent the production technology $\phi(x)$ by the following cubic polynomial in a single input variable x :

$$\phi(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3$$

The input variable x is generated from the uniform distribution over the interval [1,4]. The coefficients a_0 , a_1 , a_2 and a_3 determine the properties of the production of technology. For specificity we present the results for a set of coefficient values ($a_0 = -37$, $a_1 = 48$, $a_2 = -12$, $a_3 = 1$) that ensure that $\phi(x)$ is continuous, monotone increasing and concave in the range [1,4]. Similar results hold with other continuous, monotone increasing and concave functions as well.

Our experimental design also includes a single contextual variable z generated from the uniform distribution in the range [0,1]. We assign a value of 0.2 to β , the coefficient that

1
2
3 captures the impact of the contextual variable on productivity. We generate the noise variable, v ,
4
5 from a mean zero, two-side truncated normal distribution with upper and lower bounds at $6\sigma_v$
6
7 and $-6\sigma_v$, and report results initially for $\sigma_v = 0.04$. We generate the inefficiency variable, u , from
8
9 a half-normal distribution with parameter $\sigma_u = 0.15$. The random variables x , z , u and v are
10
11 drawn from independent probability distributions. Finally, the logarithm of the actual output y
12
13 for each draw is calculated as $\ln y = \ln(-37+48x-12x^2+x^3) - 0.2z + v - u$.
14
15
16

17
18 The choice of values for the parameters characterizing the composed error distribution is
19
20 based on guidance from simulation results from prior literature (Olson, Schmidt and Waldman
21
22 1980, Banker, Gadh and Gorr 1993, Banker, Chang and Cooper 2004) that have used data
23
24 generating processes involving composed error distributions. Based on the chosen parameter
25
26 values and the distributions underlying the composed error, the ratio of the variance of the
27
28 inefficiency term u and the variance of the measurement error v in our experiment is 5.11 which
29
30 corresponds to variance ratios for relatively moderate measurement error situations that prior
31
32 studies have used in their simulations.
33
34
35

36
37 A couple of other points are worth noting. The expected value of efficiency $E(e^{-u})$ for our
38
39 experiment is 0.89. This expected value is consistent with empirical estimates of this parameter
40
41 in previous DEA studies (Banker, Gadh and Gorr 1993). More than half of the variance (62.4%)
42
43 of the logarithm of the actual output to frontier output ratio, $\ln(y/\phi(x))$, arises from the
44
45 inefficiency term, another 12.2% from the measurement error and the remaining 25.4% from the
46
47 contextual variable component.
48
49

50 **4.2 Estimation Methods**

51
52 We employ twelve different estimation methods on the simulated data to evaluate the
53
54 impact of the contextual variable z on productivity. Five of these are one-stage parametric
55
56
57
58
59
60

procedures, five follow a two-stage parametric approach and the remaining two use DEA in the first stage followed by a parametric method in the second stage. We use ML estimation or corrected ordinary least squares (COLS) in the first stage of two-stage methods to estimate individual efficiencies. We describe below the twelve estimation methods in greater detail.

Methods 1 and 2: DEA-based Procedures

Our first two methods use DEA in the first stage followed by either OLS or MLE in the second stage. We use the standard BCC (Banker, Charnes and Cooper 1984) linear program in the first stage to estimate the output-based technical efficiency, $\hat{\theta} \leq 1$, by performing DEA on the input-output observations (x_j, y_j) , $j = 1, \dots, N$. In the DEA+OLS method, we estimate the relationship $\ln \hat{\theta} = \beta_0 - \beta z + e_1$ using OLS in the second stage yielding a consistent estimator for β . In the DEA+MLE method, we perform ML estimation in the second stage on $e_2 = \ln \hat{\theta} - \beta_0 + \beta z$ by assuming that the probability distribution of e_2 is a normal-half-normal convolution.

Method 3: One-stage, Cubic polynomial, MLE

We perform MLE in one step on the error term $e_3 = v - u = \ln(y) - \ln(a_0 + a_1x + a_2x^2 + a_3x^3) + \beta z$ assuming a normal-half-normal composed error structure for e_3 (Aigner, Lovell and Schmidt 1977). This one-shot ML estimation directly yields an estimator for β as well as estimators for a_0 , a_1 , a_2 and a_3 and the parameters underlying the composed error distribution.

Method 4: Two-stage, Cubic Polynomial, MLE

In the first step, we perform MLE on $e_4 = v - (u + \beta z) = v - u^* = \ln(y) - \ln(a_0 + a_1x + a_2x^2 + a_3x^3)$ assuming that v is distributed as normal and u^* as half-normal. Using the results from this first stage MLE, we obtain Jondrow et al.'s conditional mean estimator, $E(u^* | e_4)$, for every observation. In the second step, we perform another MLE on $e_5 = -E(u^* | e_4) - b_0 + \beta z$ assuming

that the probability distribution of e_5 is a normal-half-normal convolution. This second ML procedure yields an estimator for β .

Method 5: One-stage, Translog, OLS

We specify a translog form for the production function and incorporate the contextual variable in the set of independent variables. We then use OLS to estimate the parameters of $\ln(y)$

$$= b_0 + b_1 \ln(x) + b_2 (\ln(x))^2 - \beta z + e_6.$$

Method 6: Two-stage, Translog, COLS

We first regress $\ln(y)$ on $\ln(x)$ and $(\ln(x))^2$. Let e_7 be the residual from this regression. We next use COLS (Olson et al. 1980) to estimate the parameters of the composed error distribution from the second and third moments of e_7 . We estimate individual inefficiencies using Jondrow et al.'s conditional mean estimator $E(u^* | e_7)$. In the second stage, we estimate the relationship $-E(u^* | e_7) = c_0 - \beta z + e_8$ using OLS.

Method 7: One-stage, Translog, MLE

We specify a translog form for the production function and incorporate the contextual variable in the set of independent variables. Unlike method 5, however, we use ML instead of OLS estimation to the residual $e_9 = \ln(y) - \{b_0 + b_1 \ln(x) + b_2 (\ln(x))^2\} + \beta z$ to estimate the various parameters, assuming a normal-half-normal composed error structure for e_9 (Aigner, Lovell and Schmidt 1977).

Method 8: Two-stage, Translog, MLE

In the first step, we perform MLE on $e_{10} = v - (u + \beta z) = v - u^* = \ln(y) - \{b_0 + b_1 \ln(x) + b_2 (\ln(x))^2\}$ assuming that v is distributed as normal and u^* as half-normal. Using the results from this first stage MLE, we obtain Jondrow et al.'s conditional mean estimator, $E(u^* | e_{10})$, for every

observation. In the second step, we perform another MLE on $e_{11} = -E(u^* | e_{10}) - b_0 + \beta z$ assuming that the distribution of e_{11} is a normal-half-normal convolution.

Methods 9-12: Cobb-Douglas

We also use four other parametric methods that are simplified versions of methods 4-8, above. We specify the production function with only an intercept and the $\ln(x)$ term, excluding the $(\ln(x))^2$ term.

4.3 Performance of the Various Methods in Evaluating the Impact of Contextual Variable on Productivity

We generated 2000 sets of 200 observations (a total of 400,000 observations) for our simulation experiment. In each of the 2000 iterations, we estimated the impact of the contextual variable on productivity by applying each of the 12 estimation methods to a data set of 200 observations. We thus estimated 2000 values for β under each method. We use two performance measures: a) Mean absolute deviation % and b) Root mean squared deviation %

where mean absolute deviation (MAD) is $100 \times \left\{ \frac{1}{0.2} \left(\frac{1}{2000} \sum_{j=1}^{2000} |\hat{\beta}_j - 0.2| \right) \right\}$ and root mean squared deviation (RMSD) is $100 \times \left\{ \frac{1}{0.2} \sqrt{\frac{1}{2000} \sum_{j=1}^{2000} (\hat{\beta}_j - 0.2)^2} \right\}$.

Table 1 compares the performance of twelve estimation methods. The one-stage, cubic polynomial, MLE method (method 3) performs the best. Obviously, this is due to the fact that the assumed parametric form for the production function is identical to the production function used to generate frontier value and the assumed probability distribution for the composed error term is nearly identical to that used to generate the composed error data for the simulation. More interestingly, the two-stage estimation method DEA followed by OLS (method 1) that imposes

1
2
3 no assumption whatsoever on the parametric functional form, except that it is monotone
4 increasing and concave, or on the distribution of the error term, performs almost as well as the
5 one-stage, cubic polynomial, MLE method. Using MLE in the second stage after the first stage
6 DEA (method 2) also performs almost as well. The DEA-based methods also significantly
7 outperform both one-stage and two-stage parametric methods that assume translog or Cobb-
8 Douglas functional forms. Thus, the simulation evidence confirms that a two-stage procedure
9 that uses DEA in the first stage followed by OLS in the second stage is appropriate to evaluate
10 the impact of contextual variables on productivity.
11
12
13
14
15
16
17
18
19
20
21

22 Next, we evaluate how sensitive the performance of the DEA-based methods is to factors
23 such as sample size, level of impact of contextual variable on productivity and level of noise.
24 We generate 100 sets of 200 observations using the base case parameters, 100 sets of 400
25 observations again based on base case parameters, 100 sets of 200 observations with a higher
26 impact parameter ($\beta=0.4$), 100 sets of 200 observations with high measurement error ($\sigma_v = 0.1$),
27 and 100 sets of 200 observations with no measurement error ($\sigma_v = 0$). Table 2 displays the
28 performance of the DEA-based methods under the various scenarios.
29
30
31
32
33
34
35
36
37
38

39 Doubling the sample size from 200 to 400 reduces MAD and RMSD to substantially
40 lower levels. A comparison of the performance measures corresponding to the $\beta=0.4$ case with
41 the $\beta=0.2$ case indicates that DEA-based methods perform better when the importance of the
42 contextual variable increases relative to the pure inefficiency term and noise. However,
43 performance suffers when the relative level of measurement error increases as can be seen by
44 comparing the $\sigma_v = 0.1$ case with the base case. Finally, the absence of the measurement error
45 term does not seem to impact performance.
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

5. Conclusion

In this paper, we provide a statistical foundation for the analysis of the impact of contextual variables on productivity. We consider different DGPs and estimation methods and test procedures appropriate under each one. The point of departure is a parametrically specified DGP that assumes separability of the production function and the contextual variables. The second DGP assumes a monotone increasing and concave production function separable from a parametric function of the contextual variables. A two-stage approach comprising a DEA model followed by an ordinary least squares (or maximum likelihood estimation) model is shown to yield consistent estimators of the impact of the contextual variables for the second DGP. Our paper thus fills a gap in the DEA literature on the estimation of the impact of contextual variables on a production correspondence. It provides a formal statistical basis for the two-stage method used in earlier DEA studies.

Appendix: Proofs of Consistency of the Second Stage Estimation

Proof of Proposition 1:

We wish to show that $\text{plim } \hat{\beta} = \beta$, where $\hat{\beta}$ is the OLS estimator of $\tilde{\beta}$ in (14). If the true productivity scores $\ln \tilde{\theta}$ are available then the OLS estimation of

$$\ln \tilde{\theta} = \beta_0 - Z\beta + \delta \quad (\text{A1})$$

yields a consistent estimator of β . Here, when the DEA estimator $\ln \hat{\theta}$ is used in place of $\ln \tilde{\theta}$ we need to show that the OLS estimator of $\tilde{\beta}$ in

$$\ln \hat{\theta} = \tilde{\beta}_0 - Z\tilde{\beta} + \tilde{\delta} \quad (\text{A2})$$

also yields a consistent estimator of β . In finite samples, DEA estimators are biased (Banker 1993) since

$$\ln \hat{\theta}_j = \ln \tilde{\theta}_j + \eta_j \quad (\text{A3})$$

where $\eta_j \geq 0$ for all j . From (A1), (A2) and (A3), we have

$$\begin{aligned} \hat{\beta} &= -(Z'Z)^{-1}Z' \ln \hat{\theta} \\ &= -(Z'Z)^{-1}Z' (\ln \tilde{\theta} + \eta) \\ &= -(Z'Z)^{-1}Z' (-Z\beta + \delta + \eta) \\ &= \beta - Q^{-1}(Z'\delta/n) - \frac{1}{n}Q^{-1}Z'\eta \end{aligned} \quad (\text{A4})$$

First we show that the second term on the RHS converges in probability to zero. Since $\delta = -E(v-u) + (v-u)$ and since the random variables z_s , u and v are mutually independent, the random variables δ and z_s are mutually independent. Therefore, $\text{plim}(Z'\delta/n) = 0$ and the second term on the RHS of (A4) converges in probability to zero (Greene 1993, p. 353).

Next we show that the last term on the RHS in (A4) converges in probability to zero. From Banker (1993) we know that not only are DEA estimators consistent, but also $\text{plim}(\eta|Z) = 0$ for any Z . Therefore, $\text{plim}(Z'\eta) = 0$. Since both the second and third terms on the RHS of (A4) converge in probability to zero, $\text{plim} \hat{\beta} = \beta$.

Denote the variance of δ_j as σ^2 . Since the limiting distribution of $Z'\eta$ degenerates to a point located at the origin, the variance of the third term in (A4) converges to zero. Using this result, the limiting distribution of $\sqrt{n}(\hat{\beta} - \beta)$ can be derived as $N[\mathbf{0}, \sigma^2 \mathbf{Q}^{-1}]$ as in the standard case of OLS estimation (Greene p. 353). As a consequence, the asymptotic distribution of $\hat{\beta}$ can be derived as $N[\beta, (\sigma^2/n) \mathbf{Q}^{-1}]$.

Proof of Proposition 2:

Our proof is based on Bierens (1994) and Gstach (1998). First stage DEA provides consistent estimators $\hat{\theta}$ for the productivity variable $\tilde{\theta}$. Let the p.d.f. of the error term ε be given by $f(\varepsilon; \alpha)$, and let $\xi = (\alpha, \beta)$. Let Θ and Ξ , respectively, denote the probability space of $\tilde{\theta}$ and ξ .

Recall from (13) that

$$\varepsilon = Z\beta + V^M + \ln \tilde{\theta}$$

The estimator $\hat{\varepsilon}$ for ε , therefore, is

$$\hat{\varepsilon} = Z\hat{\beta} + V^M + \ln \hat{\theta}$$

By the assumed compactness of the input space X , its interior $\mathfrak{I}(X)$ is well defined. Similarly, the interior $\mathfrak{I}(Z)$ is also well defined for the contextual variables space Z . Let $\Omega \equiv X \times Z \times [0, 1]$ denote the probability space under consideration.

$$\text{Let } \hat{Q}_n(\xi) \equiv \ln \left[\prod_n f_{\hat{\varepsilon}}(\hat{\varepsilon} | \xi) \right] / n \text{ and } Q_n(\xi) \equiv \ln \left[\prod_n f_{\varepsilon}(\varepsilon | \xi) \right] / n$$

define, respectively, the mean log-likelihood function based on the estimator $\hat{\varepsilon}$ and the true value ε . Further, let

$$Q(\xi) \equiv \lim_{n \rightarrow \infty} E[Q_n(\xi)]$$

To prove that the ML estimator $\hat{\xi}_n$, implicitly defined by

$$\hat{Q}_n(\hat{\xi}_n) \equiv \sup_{\xi \in \Xi} \ln \left[\prod_n f_{\hat{\varepsilon}}(\hat{\varepsilon} | \xi) \right] / n,$$

converges pseudo-uniformly in probability to ξ_0 , where ξ_0 is a unique point in Ξ such that

$$Q(\xi_0) = \sup_{\xi \in \Xi} Q(\xi), \text{ it is sufficient to show that } \hat{Q}_n(\xi) \xrightarrow{\text{Prob}} Q(\xi) \text{ pseudo-uniformly}$$

$\forall \xi \in \Xi$ (Bierens 1994). We prove below that this is indeed the case and, therefore, the ML estimator of β is consistent.

Proof of $\hat{Q}_n(\xi) \xrightarrow{\text{Prob}} Q(\xi)$

The proof proceeds in two steps. First we show that $\hat{Q}_n(\xi) \xrightarrow{\text{Prob}} Q_n(\xi)$ and

$$Q_n(\xi) \xrightarrow{\text{Prob}} Q(\xi) \text{ and then use these results to show that } \hat{Q}_n(\xi) \xrightarrow{\text{Prob}} Q(\xi).$$

We prove by contradiction that $\hat{Q}_n(\xi) \xrightarrow{\text{Prob}} Q_n(\xi)$. Suppose this is not the case. Then

$$\lim_{n \rightarrow \infty} \Pr[\sup_{\xi \in \Xi} |\hat{Q}_n(\xi) - Q_n(\xi)| \leq \varepsilon_0] < 1 \text{ for some } \varepsilon_0 > 0. \text{ Because of the continuity of the supremum}$$

function this implies that there exists a set $X' \times Z' \times \Theta' \subseteq \Omega$ and an infinite dimensional vector

$$\{\mathbf{x}_\infty, \mathbf{z}_\infty, \boldsymbol{\theta}_\infty\} \equiv \{x_i, z_i, \theta_i\}_{i=1}^\infty \text{ such that } \{x_i, z_i, \theta_i\} \in X' \times Z' \times \Theta' \forall i = 1, 2, \dots \text{ and}$$

$$\sup_{\xi \in \Xi} |\hat{Q}_n(\xi) - Q_n(\xi)| > \varepsilon_0. \text{ But when } n \rightarrow \infty \text{ the difference between } \hat{\theta}_i \text{ and } \tilde{\theta}_i \text{ vanishes for all } i \text{ and,}$$

therefore, the difference between $\hat{Q}_n(\xi)$ and $Q_n(\xi)$ also vanishes for all $\xi \in \Xi$. Therefore, as

$n \rightarrow \infty$, $\sup_{\xi \in \Xi} |\hat{Q}_n(\xi) - Q_n(\xi)| = 0$ and we then have $0 > \varepsilon_0$ which contradicts the starting condition.

So, $\hat{Q}_n(\xi) \xrightarrow{Prob} Q_n(\xi)$ pseudo-uniformly. If we define $\xi_1(\mathbf{x}_n, \mathbf{z}_n, \boldsymbol{\theta}_n)$ as that value of ξ that maximizes the absolute difference $|\hat{Q}_n(\xi) - Q_n(\xi)|$, this is equivalent to saying that for all $\varepsilon_1 > 0$, there exists a set

$$\begin{aligned} \Omega_{1,n} &\equiv \left\{ \{\mathbf{x}_n, \mathbf{z}_n, \boldsymbol{\theta}_n\} : |\hat{Q}_n(\xi_1(\cdot)) - Q_n(\xi_1(\cdot))| \leq \varepsilon_1 \right\} \\ &\text{with } \lim_{n \rightarrow \infty} \Pr \left[\{\mathbf{x}_n, \mathbf{z}_n, \boldsymbol{\theta}_n\} \in \Omega_{1,n} \right] = 1. \end{aligned} \quad (\text{A5})$$

Recall that $Q_n(\xi)$, the mean log-likelihood function based on the values of the random variable $\varepsilon = v - u$, is well defined for all $\xi \in \Xi$, since the density $f(\varepsilon)$, given by (4), is continuous.

Therefore, $Q_n(\xi) \xrightarrow{Prob} Q(\xi)$ pseudo-uniformly. If we define $\xi_2(\mathbf{x}_n, \mathbf{z}_n, \boldsymbol{\theta}_n)$ as that value of ξ that maximizes the absolute difference $|Q_n(\xi) - Q(\xi)|$, this is equivalent to saying that for all $\varepsilon_2 > 0$, there exists a set

$$\begin{aligned} \Omega_{2,n} &\equiv \left\{ \{\mathbf{x}_n, \mathbf{z}_n, \boldsymbol{\theta}_n\} : |Q_n(\xi_2(\cdot)) - Q(\xi_2(\cdot))| \leq \varepsilon_2 \right\} \\ &\text{with } \lim_{n \rightarrow \infty} \Pr \left[\{\mathbf{x}_n, \mathbf{z}_n, \boldsymbol{\theta}_n\} \in \Omega_{2,n} \right] = 1 \end{aligned} \quad (\text{A6})$$

Combining (A5) and (A6), defining $\Omega_n = \Omega_{1,n} \cap \Omega_{2,n}$ and utilizing the inequality

$$|\hat{Q}_n(\xi) - Q_n(\xi)| + |Q_n(\xi) - Q(\xi)| \geq |\hat{Q}_n(\xi) - Q(\xi)| \quad (\text{A7})$$

we immediately get the following result:

$$\begin{aligned} \text{For all } \varepsilon_1, \varepsilon_2 > 0, \text{ there exists } \Omega_n &\equiv \left\{ \{\mathbf{x}_n, \mathbf{z}_n, \boldsymbol{\theta}_n\} : |\hat{Q}_n(\xi_3(\cdot)) - Q(\xi_3(\cdot))| \leq \varepsilon_1 + \varepsilon_2 \right\} \\ &\text{with } \lim_{n \rightarrow \infty} \Pr \left[\{\mathbf{x}_n, \mathbf{z}_n, \boldsymbol{\theta}_n\} \in \Omega_n \right] = 1 \end{aligned} \quad (\text{A8})$$

1
2
3 where $\xi_3(\mathbf{x}_n, \mathbf{z}_n, \boldsymbol{\theta}_n)$ is defined as that value of ξ that maximizes the absolute difference

4
5
6 $|\hat{Q}_n(\xi) - Q(\xi)|$. But (A8) can be true only if $\hat{Q}_n(\xi) \xrightarrow{Prob} Q(\xi)$ pseudo-uniformly leading to

7
8
9 the desired result.
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For Peer Review

References

- 1
2
3
4
5
6 Aigner, D.J., and S. F. Chu. 1968. On Estimating the Industry Production Function. *American*
7
8 *Economic Review*. September 1968, 826-839.
9
10
11 Aigner, D.J., Lovell, C.A.K. and Schmidt, P., 1977. Formulation and Estimation of Stochastic
12
13 Frontier Production Function Models. *Journal of Econometrics*, 6, 21-37.
14
15
16 Banker, R.D., 1993. Maximum Likelihood, Consistency and Data Envelopment Analysis: A
17
18 Statistical Foundation. *Management Science*, October, 1265-1273.
19
20
21 Banker, R.D., Chang, H. and Cooper, W.W., 2004. A Simulation Study of DEA and Parametric
22
23 Frontier Models in the Presence of Heteroscedasticity. *European Journal of Operational*
24
25 *Research*. 153, 624-640.
26
27
28 Banker, R.D., Charnes, A. and Cooper, W.W., 1984. Models for the Estimation of Technical and
29
30 Scale Inefficiencies in Data Envelopment Analysis. *Management Science*. 30, 1078-1092.
31
32
33 Banker, R.D., V. M. Gadh, and W. L. Gorr., 1993. A Monte Carlo Comparison of Two
34
35 Production Frontier Estimation Methods: Corrected Ordinary Least Square and Data
36
37 Envelopment Analysis. *European Journal of Operational Research*. 67, 332-343.
38
39
40 Banker, R.D., Janakiraman, S. and Natarajan, R., 2002. Evaluating the Adequacy of Parametric
41
42 Functional Forms in Estimating Monotone and Concave Production Functions. *Journal of*
43
44 *Productivity Analysis*. 17, 111-132.
45
46
47 Bierens, H. J., 1994. Estimation, Testing, and Specification of Cross-section and Time Series
48
49 Models. *Topics in Advanced Econometrics*, Cambridge University Press, Cambridge, Great
50
51 Britain.
52
53
54 Farrell, M.J., 1957. The Measurement of Productive Efficiency. *Journal of the Royal Statistical*
55
56 *Society (A, general)* 120, pt. 3, 253-290.
57
58
59
60

- 1
2
3 Forsund, F.R., 1999. The Evolution of DEA – The Economics Perspective. *Working Paper*,
4 University of Oslo, Norway.
5
6
7
8 Greene, W.H., 1980. Maximum Likelihood Estimation of Econometric Frontier Production
9 Functions. *Journal of Econometrics*, 13, 27-56.
10
11
12 Greene, W.H., 1993. *Econometric Analysis*. Prentice-Hall. Englewood Cliffs, New Jersey.
13
14
15 Grosskopf, S., 1996. Statistical Inference and Nonparametric Efficiency: A Selective Survey.
16
17 *Journal of Productivity Analysis*, 7, 161-176.
18
19
20 Gstach, D., 1998. Another Approach to Data Envelopment Analysis in Noisy Environments: DEA+.
21
22 *Journal of Productivity Analysis*, 9, 161-176.
23
24
25 Jondrow, J., Lovell, C.A.K., Materov, I.S. and Schmidt, P., 1982. On The Estimation of Technical
26
27 Inefficiency in the Stochastic Frontier Production Function Model. *Journal of Econometrics*, 19,
28
29 233-238.
30
31
32 Kalirajan, K.P., On Measuring the Contribution of Human Capital to Agricultural Production.
33
34 *Indian Economic Review*. 24, 247-261.
35
36
37 Meusen, W. and van den Broeck, J., 1977. Efficiency Estimation from Cobb-Douglas
38
39 Production Functions with Composed Error. *International Economic Review*. June, 435-444.
40
41
42 Olson, J. A., P. Schmidt, and D. M. Waldman., 1980. A Monte Carlo Study of Stochastic
43
44 Frontier Production Functions. *Journal of Econometrics*. 13, 67-82.
45
46
47 Pitt, M. M., and L. F. Lee., 1981. Measurement and Sources of Technical Inefficiency in the
48
49 Indonesian Weaving Industry. *Journal of Development Economics*. 9, 43-64.
50
51
52 Ray, S., 1991. Resource-use Efficiency in Public Schools: A Study of Connecticut Data.
53
54 *Management Science*. 1620-1628.
55
56
57
58
59
60

1
2
3 Schmidt, P., 1976. On the Statistical Estimation of Parametric Frontier Production Functions.
4
5 *Review of Economics and Statistics*, May, 238-239.
6

7
8 Schmidt, P., 1985. Frontier Production Functions. *Econometric Review*. 4, 289-328.
9

10 Schmidt, P., and H. Wang., 2002. One-Step and Two-Step Estimation of the Effects of
11
12 Exogenous Variables on Technical Efficiency Levels. *Journal of Productivity Analysis*, 18, 129-
13
14
15 144.
16

17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For Peer Review

Table 1

Performance Comparison of Various Estimation Methods in Evaluating Impact of Contextual Variable

Base Case: Weight on contextual variable $\beta = 0.2$, $\sigma_u = 0.15$, $\sigma_v = 0.04$, $E(e^{-u}) = 0.890729$, Sample size=200, Number of iterations =2000

Estimation Method	Mean Absolute Deviation %	Root Mean Squared Error %
Two-stage, DEA + OLS	10.5	13.2
Two-stage, DEA + MLE	11.0	14.0
One-stage, Cubic Polynomial, MLE	8.5	10.7
Two-stage, Cubic Polynomial, MLE	69.5	70.3
One-stage, Translog, OLS	25.0	32.3
Two-stage, Translog, COLS	27.0	34.5
One-stage, Translog, MLE	22.0	35.3
Two-stage, Translog, MLE	55.5	59.6
One-stage, Cobb-Douglas, OLS	42.0	54.0
Two-stage, Cobb-Douglas, COLS	42.0	53.8
One-stage, Cobb-Douglas, MLE	18.5	23.6
Two-stage, Cobb-Douglas, MLE	43.0	48.4

Table 2**Performance Comparison of DEA-Based Procedures under Different Scenarios**

Base case: Weight on contextual variable $\beta = 0.2$, $\sigma_u = 0.15$, $\sigma_v = 0.04$, $E(e^{-u}) = 0.890729$, Sample size=200, Number of iterations =100

Scenario	Estimation Method	Mean Absolute Deviation %	Root Mean Squared Error %
Base Case	DEA+OLS	10.5	12.9
Base Case	DEA+MLE	11.0	14.4
Larger Sample (N=400)	DEA+OLS	2.0	7.5
Larger Sample (N=400)	DEA+MLE	4.0	7.5
Larger β ($\beta = 0.4$)	DEA+OLS	5.5	6.9
Larger β ($\beta = 0.4$)	DEA+MLE	5.5	7.1
High Noise ($\sigma_v = 0.10$)	DEA+OLS	14.0	17.2
High Noise ($\sigma_v = 0.10$)	DEA+MLE	14.5	17.6
No Noise ($\sigma_v = 0$)	DEA+OLS	10.5	12.6
No Noise ($\sigma_v = 0$)	DEA+MLE	11.5	14.1