Technical Efficiency Estimation with Multiple Inputs and Multiple Outputs Using Regression Analysis

Abstract

Regression and linear programming provide the basis for popular techniques for estimating technical efficiency. Regression based approaches are typically parametric and can be both deterministic or stochastic where the later allows for measurement error. In contrast, linear programming models are nonparametric and allow multiple inputs and outputs. The purported disadvantage of the regression based models is the inability to allow multiple outputs without additional data on input prices. In this paper, deterministic cross-sectional and stochastic panel data regression models that allow multiple inputs and outputs are developed. Notably, technical efficiency can be estimated using regression models characterized by multiple-input, multiple-output environments without input price data. We provide multiple examples including a Monte Carlo Analysis.

Keywords: DEA, stochastic frontier analysis, joint production
1. Introduction

In the economics and operations research literature there are two distinct approaches for estimating technical efficiency. Both regression and linear programming techniques have been employed to measure performance relative to an estimated frontier. The starting point for both literatures is Farrell (1957), which provided a conceptual framework for efficiency measurement. Farrell showed that technical and allocative inefficiency could be measured relative to the observed isoquants with equi-proportional measures. Farrell also illustrated efficiency using piecewise linear isoquants.

Aigner and Chu (1968) extended Farrell’s work by applying programming models to parametric functional forms to measure production in deterministic models where all deviations from the frontier are one-sided and due to inefficiency. Winsten (1957) suggested and Greene (1980) showed that OLS could be used to estimate inefficiency relative to a frontier with one-sided deviations. Since the parameters of the production function are estimated consistently, one need only correct the intercept term by adding the largest residual to the intercept in a production environment. This technique is referred to as Corrected OLS (CLOS).

CLOS is limited due to the nature of regression analysis; only one output is allowed in the production function. Lovell et al. (1994) proposed a solution in the multiple output case by specifying a distance function, exploiting homogeneity and rearranging terms to specify the production process with one output used as the dependent variable while treating all other outputs as independent variables. This method called the stochastic distance function (SDF) has been popularized by Grosskopf et al. (1997), Coelli and Perelman (1999) and Coelli and Perelman (2000). The asymmetric
treatment of a single output in SDF has been criticized by Atkinson and Primont (2002) for creating an endogeneity problem, see also Vinod (1969).

Further, the estimated output isoquants often do not satisfy the concavity or quasi-concavity properties implied by production theory, Sauser et al. (2006). While O’Donnell and Coelli (2005) attempt to remedy this issue, they need to place additional structure on the problem by generating additional data and specifying \textit{a priori} the median efficiency level. Additionally, the remedy also suffers from the asymmetric treatment of outputs. Even though these considerable limitations exist, the stochastic distance function approach has been widely used in the production literature; see for example Atkinson et al. (2003), Fernandez et al. (2005), Smith and Street (2005), and Kumbhakar et al. (2007), which use the standard stochastic distance function and Yan et al. (2009) and Feng and Serletis (2010), which use the O’Donnell and Coelli (2005) remedy. As pointed out by Coelli and Perelman (2000), a maintained advantage of the SDF approach is the ability to estimate non-separable production. However, regression based approaches have other advantages such as providing goodness-of-fit and other statistics that help evaluate the overall model. One of the contributions of this paper is a new approach that extends COLS to handle multiple outputs. Unlike the SDF approach, our model treats outputs symmetrically and satisfies proper curvature in output space without strong \textit{a priori} assumptions.

Another limitation of the COLS model (and all other deterministic models) is the inability to properly account for measurement error. From an econometrician’s view, attributing all deviations to inefficiency in production is not appealing. Instead, deviations from the frontier can occur not only from inefficient behavior but also from
measurement error and statistical noise. A large body of research developed based on the pioneering work of Aigner et al. (1977) and Meeusen and van den Broeck (1977), both of which laid the foundation for the stochastic frontier approach (SFA). These papers assumed that the deviation from the frontier consisted of an overall error composed of inefficiency and statistical noise.

Jondrow et al. (1982) provided the means to measure observation specific inefficiency based on the expected value of inefficiency conditional on the observed overall inefficiency. The conceptual treatment is well-justified. However, as shown by Ruggiero (1999) and Ondrich and Ruggiero (2001), the cross-sectional models do not hold any advantages over deterministic models; the expected value of inefficiency given the observed overall error is perfectly correlated with the overall error itself. Using distributional assumptions to derive efficiency estimates does not allow noise to effect the ranking of observations.

Schmidt and Sickles (1984) overcome the problem of assuming a distributional form for the inefficiency component by extending the SFA model using a fixed effects panel data model. If one assumes that all variation other than inefficiency is controlled for by observable variables, then the individual specific fixed effect is composed of each individual’s inefficiency. The main drawback to this approach is that it assumes time invariant efficiency. However, this assumption may hold true for many panel data sets that include only a few years or measure data at frequent increments such as monthly or weekly. Pitt and Lee (1981) proposed a random effects panel data model—similar to the fixed effects model—that can also be used to estimate efficiency. However, it requires that the individual specific component to be uncorrelated with all control variables and
the error component. It is difficult to justify these assumptions in most real world situations. Additional panel models have been proposed to measure technical efficiency; however, most of them suffer from the same limitation as the cross-sectional models—they require a distributional assumption for the inefficiency component. See Battese and Coelli (1995) for a further discussion. The panel data models are effectively able to separate the effects of noise and inefficiency creating a distinct advantage over the cross-sectional SFA model which cannot separate these components effectively or DEA which assumes no noise in the data.

The alternative to the regression based approaches is Data Envelopment Analysis (DEA), a nonparametric programming model that allows multiple outputs and inputs. Popularized by Charnes et al. (1978) and later extended by Banker et al. (1984), the approach has become widely used in analyzing technical efficiency of public sector units. See also Färe and Lovell (1978). There are two main advantages that DEA has over regression based approaches. First, the technique is nonparametric in the sense that a priori specification of the production function is not required. Rather, the approach estimates the frontier using the minimum extrapolation principle under the maintained axioms of monotonicity and convexity of the production possibility set, Banker et al. (1984), although this interpretation has been challenged, see for example Chang and Guh (1991). Second, and perhaps more important, DEA easily handles multiple inputs and multiple outputs and allows direct comparisons of production possibilities without requiring additional input price data.

There have been many studies that have analyzed the performance of DEA and regression based approaches. Typically, simulation analysis is employed with a data
generating process involving a production function with only one output. Gong and Sickles (1992) compared DEA and the stochastic frontier approach with multiple outputs but relied on input prices. Banker et al. (1993) analyzed the performance of DEA relative to COLS using cross-sectional simulated data. The results indicated that COLS did not properly adjust for measurement error, DEA performed at least as well, and that both models performed worse as measurement error increased. Ruggiero (1999) used cross-sectional simulated data and showed that the stochastic frontier model does not control for measurement error and deterministic COLS performed as well. Coelli and Perelman (1999) provided a comparative analysis of a multiple-output and multiple-input technology using DEA and regression analysis implemented using SDF and found that SDF worked well.

In contrast, our approach applies a DEA based method in a first stage to provide a measure of aggregate output which is then incorporated into a second-stage regression. McDonald (2008) argues that while Tobit estimation is inappropriate, OLS provides consistent estimates in the second-stage. The primary advantage of the approach developed in this paper is the use of a nonparametric output aggregate that conforms to desirable properties of the output set and treats outputs symmetrically.

The purpose of this paper is to extend the regression based approaches to measure efficiency in multiple-input and multiple output technologies. The rest of the paper is organized as follows. In the next section, the mathematical foundations of the technology are developed. Section 3 extends COLS to multiple output technologies and section 4 extends the stochastic frontier models. Section 5 contains a Monte Carlo analysis for comparative purposes. The last section concludes with directions for further research.
2. Description of the Technology

Assume that each of $n$ DMUs employ a vector $x$ of $s$ inputs to produce a vector $y$ of $m$ outputs according to the technology $T = \{(x, y) : x \in \mathbb{R}^s_+, y \in \mathbb{R}^m_+, x \text{ can produce } y\}$. For our purposes, we define the output set as $P(x) = \{y : (x, y) \in T\}$. The standard properties on $P(x)$ discussed in Färe et al. (1994) are assumed. Following Färe et al. (1994), define the isoquant $Isoq P(x)$ as:

$$Isoq P(x) = \{y \in P(x)) : \theta y \in P(x) \text{ for } \theta > 1\}.$$  

This boundary is used to compare observed production possibilities to the boundary of the output set. DEA uses a piecewise linear approximation to the estimation of the output set (and the input set). Färe et al. (1994), prove that the piecewise linear technology $P(x)$ is closed and bounded, sufficient conditions for the existence of the efficiency measure.

The Banker et al. (1984) output-oriented DEA model to evaluate the technical efficiency of DMU “o” under the assumption of variable returns to scale (VRS) is given by:

$$F_o (x_o, y_o) = \text{Max } \theta_o$$  

s.t.

$$\sum_{j=1}^{N} \lambda_j x_{j}^o \leq x_{i}^o \quad \forall \; i = 1, ..., s$$  

$$\sum_{j=1}^{N} \lambda_j x_{j} \leq x_{i} \quad \forall \; l = 1, ..., m$$  

$$\sum_{j=1}^{N} \lambda_j = 1$$  

$$\lambda_j \geq 0 \quad \forall \; j.$$  

The general set up is shown in Figure 1, where two output sets are shown. $P(x) \subseteq P(x^1)$ with $x^1 \geq x$. Five observed production possibilities $A, B, C, D$ and $F$ are shown. It is
assumed that \( A, C, F \in P(x) \) but \( A, C, F \notin P(x') \). Production possibilities \( A, B \) and \( D \) are technically efficient with \( A \in Isoq \ P(x) \) and \( B \in Isoq \ P(x') \). Production possibilities \( C \) and \( F \) however, are technically inefficient. Based on the definition of output-oriented efficiency and the solution of (1), we have

\[
F_a(x_c, y_c) = y_{1c} / y_{1c} \quad \text{and} \quad F_a(x_d, y_d) = F_a(x_d, y_d) = 1.
\]

As mentioned in the introduction, the purported advantage of DEA over regression based approaches is the ability to estimate the production technology characterized by multiple inputs and multiple outputs without relying on input prices. Regression based approaches can measure technical efficiency assuming only one input or cost efficiency if input price data are available. Before providing a regression based approach that allows multiple outputs in a deterministic model, we define the aggregate output set.

**Definition:** \( P_A = \bigcup_{j=1}^{N} P(x_j) \) is the aggregate output set.

Because the aggregate output set is the union of a finite number of compact sets, it too must be compact. As a result, \( P_A \) is closed and bounded thus guaranteeing the existence of a distance function from any element in \( P_A \) to the boundary of \( P_A \). The boundary is defined by the isoquant:

\[
Isoq \ P_A = \{ y \in P_A : \theta y \notin P_A \ \text{for} \ \theta > 1 \}.
\]

Furthermore, we can appeal to a piecewise linear approximation to generate \( Isoq \ P_A \).

Given the assumptions on each output set \( P(x) \), the aggregate output set \( P_A \) can be thought of as the output set associated with the highest aggregate input level. Of note, the relevant properties (no free lunch, output disposability, boundedness and convexity) on
the production technology discussed in Färe et al. (1994) hold for aggregate output set $P_A$.

The linear programming model to measure the distance $F_A$ for DMU “o” to the aggregate output set is given by:

$$F_A(y_o) = \text{Max } \Theta_o$$

$$\text{s.t.}$$

$$\sum_{j=1}^{N} \Lambda_j y_{kj} \geq \Theta_o y_{ko}, \quad \forall k = 1, \ldots, s$$

$$\Lambda_j \geq 0, \quad \forall j.$$

(2)

Note that this model is similar to the output-oriented DEA model assuming variable returns to scale with the exclusion of the input constraints. Model (2) produces an estimated output isoquant $Isoq P_A$. This model has been used previously to compare observations based strictly on their multi-criteria output vector; in this case, the input constraints can be dropped only if the convexity constraint is included (see Lovell and Pastor, 1999). In this paper (2) is used to aggregate outputs; because separability is assumed, it is not necessary to consider inputs in this first stage. The output aggregate proposed is a measure of output relative to the estimated isoquant $Isoq P_A$.

Returning to Figure 1, the solution to (2) leads to $F_A(x_B, y_B) = F_A(x_D, y_D) = 1$, $F_A(x_C, y_C) = y_{1B} / y_{1C}$, $F_A(x_A, y_A) = y_{1B} / y_{1A}$ and $F_A(x_F, y_F) = y_{1D} / y_{1F}$. Production unit $F$ poses a special problem; the output constraint for $y_2$ does not hold with equality; excess slack exists after radial projection leading to a shadow price of zero. This is a well-known problem in the DEA literature. However, more recently, Johnson and Ruggiero (2009) show that the Farrell measure adequately measures performance even in the
presence of slack. The measures can be decomposed into products of efficiency and distances between isoquants. For example,

\[ F_A(x_c, y_c) = \frac{y_{1b}}{y_{1c}} = F_o(x_c, y_c) \times F_A(x_c, y_c). \]

This distance function captures inefficiency (comparing \( C \) to \( A \)) and the distance between frontiers (comparing \( A \) to \( B \)).

Production units farther from the aggregate output set produce lower output aggregates; hence \( S = F_A^{-1} = 1/F_A \) provides an index of aggregate observed output. This measure can be used in a second-stage regression where aggregate production is regressed on observed inputs. This second stage approach, like all regression-based models, requires \textit{a priori} specification of the production function. However, a translog model can be used for flexibility. In the next section, we discuss estimating production and efficiency and provide some illustrative examples.

3. Estimation of Multiple Output Production Using Regression

Regression analysis begins by specifying a production function. We estimate a multiple input, multiple output, production function in a cross-sectional analysis

\[ h(y_i) = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, N \]

(3)

Where \( y_i \) is the output vector for the \( i^{th} \) firm, \( x_i \) is the input vector for the \( i^{th} \) firm, \( f \) is an input aggregate function, \( h \) is an output aggregate function, and \( \varepsilon_i = v_i - u_i \) is a composite error term that captures all deviations from the production frontier. In this paper we make the axiomatic assumption that the production function is separable.

To describe the data generation process in more detail, \( v_i \) is a random disturbance term that includes the effects of omitted factors, measurement errors, and other stochastic noise. Assume \( v \) is a truncated normal variable with zero-mean and \( f_v \) is a probability density distribution consistent with that specification, see Gstach (1998) and Banker and
Natarajan (2008) for further examples of productivity analysis in the presence of a truncated noise term. Also, $u_i \geq 0$ is random inefficiency of firm $i$. We assume the existence of well-behaved probability density functions $f_u$ with left-truncation at zero. Variables $v_i$ and $u_i$ are assumed to be independently distributed random variables that are uncorrelated with the input variables $x_i$, and with each other. Assume that variables $x$ are randomly sampled from domain $D_x$. Further, the joint density of the random model variables is denoted as $f_d(x,u,v)$.

A desirable property of any estimator is consistency. Thus for (2) we show it consistently estimates $P_A$:

**Theorem 1:** If the following five assumptions are satisfied:

1. The boundary of $T$ is a monotonic and concave function in $x$,
2. The underlying production function, $h(y_i) = f(x_i)$, is separable,
3. sequence $\{y_i, x_i\}, i=1,\ldots,n$ is a random sample of independent observations,
4. noise terms $v_i$ have a truncated distribution: $|v| \leq V^M 1$, $f_v(V^M) > 0$,
5. the joint density $f_d$ satisfies $f_d(x,0,V^M) > 0 \ \forall x \in D_x$,

then the estimator (2) is a consistent estimator for the boundary of $P_A$, in the following sense

$$\lim_{n \to \infty} Isoq(x_i) = Isoq(P_A) + V^M \ \text{for all} \ i = 1,\ldots,n.$$

**Proof:** See appendix.

Given that (2) is a consistent estimator, with a sufficiently large sample our measure of aggregate output $S = F_A^{-1}$ can be used in a subsequent regression. Further we verify this results through several examples. The first two examples we consider are deterministic and assume no measurement error. For both examples, we adopt the technology used in Färe *et al.* (1994). In particular, technology is represented by a two-input, two-output
transformation function with a Constant Elasticity of Transformation (CET) output aggregate and a Cobb-Douglas input aggregate. We use OLS to estimate the model and use COLS to estimate technical efficiency (example 2). COLS is used in the second stage and given that production is separable and the first stage is estimated consistently as shown above, then the second stage estimation can be shown to be consistent as in Greene (1980).

**Example 1: Färe, Grosskopf and Lovell Data**

After applying model (2) above and obtaining our measure of aggregate output $S$ we apply OLS to estimate production. Here we use the same data set generated in Färe et al. (1994); these data are reported in Table 1. Efficient production is given by the function $h(y) = f(x)$, where

$$h(y) = (0.5y_1^2 + 0.5y_2^2)^{0.5}$$

and

$$f(x) = (x_1^{0.5} x_2^{0.5} )^\delta.$$  

The parameter $\delta$ was used to allow variable returns to scale. We note that $\varepsilon = 0$, leading to efficient production without measurement error.

Four different values for $f(x)$ and $h(y)$ were assumed with $\delta$ taking on values of 0.898, 1.0 and 0.927. Applying the technique from Section 2, we obtain an estimate $S$ of the output aggregate; these results are also reported in Table 1. As shown, $S$ is a good index of aggregate output $h(y)$ where $S$ is approximately equal to $h(y)$ divided by the maximum $h(y)$. The correlation between $S$ and $h(y)$ is 0.999.

Given our estimate $S$ of $h(y)$, the next step is to estimate the relationship between $S$ and the inputs. Due to variable returns to scale, a translog equation was estimated using
OLS. The regression results are presented in Table 2 and the resulting predicted value $\hat{S}$ is included in Table 1. All parameters are statistically significant at the 1 percent level and the resulting $R^2$ is approximately 1. These results are obtained with a small sample size of 20. The correlation between $h(y)$ and $\hat{S}$ is 1.00. Importantly, the results suggest that regression can be used to estimate multiple-output and multiple-input production relationships. While informative, this example is limited for our purposes because all observations are assumed to be efficient. In the next example, we allow inefficiency and perform a comparative analysis of DEA and the multiple output COLS model.

Example 2: Multiple Inputs, Multiple Outputs under CRS

In this example, we assume a Cobb-Douglas input aggregate $f(x) = x_1^{0.4}x_2^{0.6}$. This input aggregate is similar to the one used in Färe et al. (1994) with $\delta = 1$ imposing constant returns to scale. Input data were generated for 100 DMUs with $x_1, x_2 \sim N(100,25)$. Further, inefficient behavior is allowed where $g(x) = e^{-u}f(x)$, and $u \sim N(0,0.2)$. Three additional extraneous inputs (labeled $x_3$, $x_4$ and $x_5$) were generated using the same distribution as the appropriate inputs. Inappropriate inclusion of these irrelevant inputs will allow sensitivity of the estimators to model misspecification.

Two output variables, $y_1$ and $y_2$, are generated using the following procedure. Two random variables $z_1$, $z_2$ were generated assuming $z_1, z_2 \sim N(60,10)$. Using the output aggregate $h(y) = (0.5y_1^2 + 0.5y_2^2)^{0.5}$ recommended by Färe et al. (1994), we construct $h(z) = (0.5z_1^2 + 0.5z_2^2)^{0.5}$. Variables $z_1$ and $z_2$ are scaled by $\gamma = \sqrt{\frac{2g(X)^2}{z_1^2 + z_2^2}}$ to obtain observed outputs $y_1 = \gamma z_1$ and $y_2 = \gamma z_2$. 
We obtain \( h(y) = (0.5y_1^3 + 0.5y_2^3)^{0.5} = e^{-u} f(x). \) Descriptive statistics of the observed inputs and outputs are provided in Table 3. Given that constant returns to scale were assumed, the CCR DEA model, i.e., model (1) without the convexity constraint, is used. The DEA model is applied to four scenarios depending on variable selection. The first scenario is the correctly specified model based on the data generating process. Three other models are considered; all scenarios include the appropriate inputs \( x_1, x_2 \) while scenarios 2 - 4 also include the extraneous “inputs”. The scenarios considered are:

Scenario 1: correct specification;

Scenario 2: incorrectly specified with inputs \( x_1 - x_3; \)

Scenario 3: incorrectly specified with inputs \( x_1 - x_4; \) and

Scenario 4: incorrectly specified with inputs \( x_1 - x_5. \)

This analysis will allow a sensitivity analysis with respect to variable selection. It is expected that DEA will perform well under scenario 1 with performance declining as the model becomes increasingly mis-specified.

In addition to DEA, our COLS approach is applied with an output aggregate obtained via solution to model (2). In this case, model (2) is used once for each DMU. The different scenarios considered require a separate regression but not generation of the output aggregate. Given the data generating process for the input aggregate, a Cobb-Douglas regression model is considered in the second stage. The regression results are reported in Table 4.

The results indicate that the two-stage model performs well in estimating the production process. The only parameters that are statistically significant are the coefficients on the correctly specified variables. All other slope parameters are
statistically insignificant. The consistent $R^2$ of approximately 0.80 indicates that unobserved inefficiency accounts for approximately 20 percent of the variation in aggregate output. Given these results, it is expected that the multiple output COLS model developed here will perform well under all scenarios. Notably, regression weights the independent variables; inclusion of extra variables that are uncorrelated with the appropriate independent variables and the output aggregate should result in statistically insignificant parameter estimates.

Per the suggestion of an anonymous reviewer, we also consider the SDF popularized by Coelli and Perelman (1999, 2000). Consistent with their approach, we assume a distance function and estimate the flexible translog functional form. Given that the generating process assumes separability, we implement their approach for enforcing separability. Three criteria are used for evaluating the methods: the mean absolute difference (MAD), the correlation and the rank correlation between true efficiency and estimated efficiency. A lower MAD indicates that the estimate is closer on average to the true efficiency. The correlation and rank correlation coefficients provides evidence of the strength of association between true and estimated efficiency. The results of the simulation analysis are reported in Table 5.

The results indicate that our multiple output COLS approach developed in this paper performs better than both DEA and the SDF in this example. Notably, the multiple output COLS measure achieves lower MADs and higher correlation and rank correlation coefficients than both DEA and the SDF. For all performance measures across all scenarios, our multiple output COLS approach outperforms both other multiple output approaches. Interestingly, the SDF achieves lower MADs than DEA but DEA achieves
higher correlations and rank correlations across scenarios. In scenario 1, where the models are correctly specified, our COLS approach has a slightly higher correlation and rank correlation coefficient and a MAD that is more than half that of DEA. As more irrelevant variables are added to the analysis, the performance of all methods declines; however, the decline is notably worse for DEA and the SDF. In the case of DEA, the MAD increases from 0.048 in scenario 1 to 0.082 in scenario 4; the correlation drops from 0.96 to 0.86 and the rank correlation decreases from 0.93 to 0.87. The results for the SDF are similar. The COLS approach, on the other hand, maintains a MAD below 0.027 and correlations above 0.925 under all scenarios. Notably, the results for our multiple output COLS are similar for scenarios 3 and 4.

4. **Multiple Outputs and the Stochastic Frontier**

In order to extend our approach to measure efficiency in the stochastic case, we now represent the technology by a transformation function

\[ h(y_{it}) = f(x_{it}, u_t, v_{it}), \]  

(4)

where \( v_t \) represents measurement error and other statistical noise and \( u \) measures firm specific, time invariant inefficiency. Unlike the inefficiency term, measurement error is allowed to vary across time. Applying (2) to our simulated data with a technology represented by (4), we obtain an observed aggregate measure of output \( S \) that is contaminated by measurement error and inefficiency.

We use a fixed effects panel data model (see Schmidt and Sickles (1984)) and a random effects panel data model to estimate the efficiency of each DMU. Maximum likelihood models must assume a parametric distribution for the inefficiency term (usually half-normal or exponential). Assuming a Cobb-Douglas functional form and
including subscripts for observation and time, the fixed effects panel data model can be written as
\[ S_{it} = \alpha + \beta x_{it} - u_i + v_{it}, \]  
(5)
where all variables are defined as before. Consistent with the interpretation of \( u_i \) as an inefficiency term, it is assumed that \( u_i > 0 \) for all \( i \). Grouping the intercept and the technical inefficiency term, equation (4) may be re-written as
\[ S_{it} = (\alpha - u_i) + \beta x_{it} + v_{it} \]
\[ = \alpha_i + \beta x_{it} + v_{it} \]  
(6)

Given the above assumption concerning the error term, equation (6) may be estimated using the standard fixed effects (‘within’) estimator. Estimates of \( u_i \) that are strictly non-negative are then given by the deviation between each DMU-specific intercept and the maximum intercept:
\[ \hat{u}_i = \max_j \{ \hat{\alpha}_j \} - \hat{\alpha}_i. \]  
(7)

The technical efficiency measure is defined as \( \exp(-\hat{u}_i) \), which is bound by zero and unity. By construction, the DMU with the highest individual intercept is deemed technically efficient. To assure the input aggregate is estimated consistently we simply recognize that production is assumed to be separable and \( S_i \) is estimated consistently as shown above. Then the arguments presented in Schmidt and Sickles (1984) regarding consistency can be used directly.

The random effects panel data model takes the same form as equation (6); however, additional assumptions are made. The random effects model assumes that \( u \) is a random variable and uncorrelated with the input variables and \( v \). This model is
estimated using a standard two-stage generalized least squares approach. Once we obtain estimates of the DMU specific random effect, we can transform it into a measure of technical efficiency just as we did with the fixed effects model; the resulting estimator of technical efficiency is consistent (see Cornwell et al. (1990)). See Kumbhakar and Lovell (2000) for more on estimating efficiency with fixed and random effects models.

5. Stochastic Frontier Monte Carlo Analysis

The starting point for our simulated analysis is the specification of production function. In order to interpret the performance of our two-stage approach, we consider first a baseline case where one output is produced:

\[
Ln \ y_{it} = 0.4 Ln \ x_{1it} + 0.6 Ln \ x_{2it} - u_i + v_{it},
\]

(8)

where constant returns to scale prevail and individual specific inefficiency \( u_i \) does not vary across time. Data were generated randomly from the following distributions:

\[
x_s \sim N(100,25), \quad s = 1,2
\]

\[
u \sim N(0,0.2)\]

\[
v \sim N(0,\sigma_v).
\]

We note that the resulting \( \sigma_u = 0.12 \). For our experiments, we fix the number of observations at 200 and the number of time periods at 10 and allow \( \sigma_v \) to take on values of 0.15, 0.2 and 0.25. In all cases, we have \( \sigma_v > \sigma_u \); the ratio \( \frac{\sigma_v^2}{\sigma_u^2} \) of measurement error variance to inefficiency variance varies from 1.56 to 4.34. Three measures are used to evaluate the performance of the estimators: the correlation, rank correlation and mean absolute deviation (MAD) between true and estimated efficiency.
For the case of the single output, we estimate technical efficiency using both random and fixed effects models. Since the data generating process is consistent with the random effects specification, we expect that the random effects model will perform better. However, because the fixed effects model provides consistent estimates, the improvement should be minimal. We replicate this process 100 times. Summary results for the random effects model are reported in Table 6. Fixed effects results are reported in Table 7. The results are as expected. The average correlation between true and estimated efficiency is slightly higher for the random effects model while the rank correlation results are nearly identical. Interestingly, the fixed effects model performed better with respect to MAD criteria than the random effects model while performing similarly on the basis of the other criteria.

The extension to the multiple output case required specification of the output aggregate. We assumed a constant elasticity of transformation output aggregate:

\[ g(y_{it}) = \left( 0.5y_{1it}^\rho + 0.5y_{2it}^\rho \right)^{1/\rho}. \]  

(9)

Here, for our application, we choose \( \rho = 2.5 \). Data for the outputs were generated as follows:

\[ y_m \sim N(100, 25), \quad m = 1, 2. \]

Similar to the procedure used in the COLS example 2 above, the generated outputs were scaled by

\[ \delta_{it} = \frac{2x_{1it}^{0.4} x_{2it}^{0.4}}{y_{1it}^{2.5} + 0.5y_{2it}^{2.5}} \]

to ensure that

\[ Ln g(\delta_{it}, y_{it}) = 0.4Ln x_{1it} + 0.6Ln x_{2it} - Ln u_i + Ln v_{it}. \]

(10)

Data generation for all variables other than output followed the same distribution used in the one output case. Given the output data, we employed linear program (2) for each
time period to obtain an index of aggregate output. The index was then used in a second stage panel model using both fixed and random effects. Average results for 100 replications are reported in Tables 6 and 7.

The results of the analysis are encouraging. The performance results found between the fixed effects and random effects models in the one output model hold true in the two output model. In addition, while the correlation and rank correlation results are lower on average, the difference is less than 0.01 in all cases. In addition, the average MAD across replications were nearly identical.

6. Conclusions

One of the main advantages of DEA over regression based approaches is the ability to handle multiple inputs and multiple outputs. In this paper, a new regression based approach was developed that overcomes this limitation. In particular, a two-stage model was developed that employs a modified DEA model to estimate the output aggregate, which is then used in regression to measure efficiency. Notably, the output aggregate is obtained via a nonparametric specification. Returns to scale assumptions are then incorporated in the second-stage regression where a translog model allows variable returns to scale. The model was tested against DEA using a variable returns to scale model using data published in Färe et al. (1994) and a simulation where constant returns to scale prevailed. The results of the simulation show that regression based approaches can be used to measure efficiency in multiple output / multiple input deterministic production environments without additional information on input prices. In the simulation example, the results illustrate that our multiple output COLS approach outperforms DEA and the SDF approach.
We also introduced a new two-stage approach for measuring technical efficiency for multiple input and output production technologies in the presence of measurement error. In the first stage, a modified DEA model was employed to obtain a measure of observed aggregate output. The resulting index is then incorporated into a second stage stochastic frontier model. We allow flexibility in the second-stage by employing either random effects or fixed effects. The models were tested using Monte Carlo analysis; the results indicate that this approach works as well as its single output counterpart. This contribution is important because one of the purported disadvantages of using the stochastic frontier approach is its inability to handle multiple outputs.
Figure 1: Representation of Technology
Table 1: Färe, Grosskopf and Lovell Example Data

<table>
<thead>
<tr>
<th>DMU</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$y_1$</th>
<th>$y_2$</th>
<th>$h(y)$</th>
<th>$f(x)$</th>
<th>$\delta$</th>
<th>$S$</th>
<th>$\hat{S}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>36.00</td>
<td>36.00</td>
<td>25.00</td>
<td>25.00</td>
<td>25</td>
<td>36</td>
<td>0.898</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>2</td>
<td>28.80</td>
<td>45.00</td>
<td>30.00</td>
<td>18.71</td>
<td>25</td>
<td>36</td>
<td>0.898</td>
<td>0.26</td>
<td>0.25</td>
</tr>
<tr>
<td>3</td>
<td>21.60</td>
<td>60.00</td>
<td>20.00</td>
<td>29.15</td>
<td>25</td>
<td>36</td>
<td>0.898</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>4</td>
<td>19.94</td>
<td>65.00</td>
<td>15.00</td>
<td>32.02</td>
<td>25</td>
<td>36</td>
<td>0.898</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>5</td>
<td>43.20</td>
<td>30.00</td>
<td>12.00</td>
<td>33.26</td>
<td>25</td>
<td>36</td>
<td>0.898</td>
<td>0.26</td>
<td>0.26</td>
</tr>
<tr>
<td>6</td>
<td>50.00</td>
<td>50.00</td>
<td>50.00</td>
<td>50.00</td>
<td>50</td>
<td>50</td>
<td>1.000</td>
<td>0.50</td>
<td>0.51</td>
</tr>
<tr>
<td>7</td>
<td>45.00</td>
<td>55.56</td>
<td>40.00</td>
<td>58.31</td>
<td>50</td>
<td>50</td>
<td>1.000</td>
<td>0.50</td>
<td>0.51</td>
</tr>
<tr>
<td>8</td>
<td>60.00</td>
<td>41.67</td>
<td>60.00</td>
<td>37.42</td>
<td>50</td>
<td>50</td>
<td>1.000</td>
<td>0.52</td>
<td>0.52</td>
</tr>
<tr>
<td>9</td>
<td>30.00</td>
<td>83.33</td>
<td>30.00</td>
<td>64.03</td>
<td>50</td>
<td>50</td>
<td>1.000</td>
<td>0.50</td>
<td>0.51</td>
</tr>
<tr>
<td>10</td>
<td>70.00</td>
<td>35.71</td>
<td>20.00</td>
<td>67.82</td>
<td>50</td>
<td>50</td>
<td>1.000</td>
<td>0.53</td>
<td>0.52</td>
</tr>
<tr>
<td>11</td>
<td>75.00</td>
<td>75.00</td>
<td>75.00</td>
<td>75.00</td>
<td>75</td>
<td>75</td>
<td>1.000</td>
<td>0.75</td>
<td>0.76</td>
</tr>
<tr>
<td>12</td>
<td>45.00</td>
<td>125.00</td>
<td>30.00</td>
<td>101.73</td>
<td>75</td>
<td>75</td>
<td>1.000</td>
<td>0.79</td>
<td>0.77</td>
</tr>
<tr>
<td>13</td>
<td>60.00</td>
<td>93.75</td>
<td>50.00</td>
<td>93.54</td>
<td>75</td>
<td>75</td>
<td>1.000</td>
<td>0.75</td>
<td>0.76</td>
</tr>
<tr>
<td>14</td>
<td>105.00</td>
<td>53.57</td>
<td>80.00</td>
<td>69.64</td>
<td>75</td>
<td>75</td>
<td>1.000</td>
<td>0.75</td>
<td>0.77</td>
</tr>
<tr>
<td>15</td>
<td>85.00</td>
<td>66.18</td>
<td>90.00</td>
<td>56.12</td>
<td>75</td>
<td>75</td>
<td>1.000</td>
<td>0.78</td>
<td>0.76</td>
</tr>
<tr>
<td>16</td>
<td>144.00</td>
<td>144.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100</td>
<td>144</td>
<td>0.927</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>17</td>
<td>115.20</td>
<td>180.00</td>
<td>115.00</td>
<td>82.31</td>
<td>100</td>
<td>144</td>
<td>0.927</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>18</td>
<td>86.40</td>
<td>240.00</td>
<td>80.00</td>
<td>116.62</td>
<td>100</td>
<td>144</td>
<td>0.927</td>
<td>1.00</td>
<td>1.01</td>
</tr>
<tr>
<td>19</td>
<td>172.80</td>
<td>120.00</td>
<td>65.00</td>
<td>125.60</td>
<td>100</td>
<td>144</td>
<td>0.927</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>20</td>
<td>201.59</td>
<td>102.86</td>
<td>60.00</td>
<td>128.06</td>
<td>100</td>
<td>144</td>
<td>0.927</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Mean: 74.73, St. Dev.: 50.66, Min.: 19.94, Max.: 201.59

Data are taken from Färe, Grosskopf and Lovell (1994). Calculations of the output aggregate and the estimated output aggregate are by the authors.
## Table 2: Example 1 Regression Results

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-5.89</td>
<td>(0.27)</td>
</tr>
<tr>
<td>$\ln x_1$</td>
<td>1.34</td>
<td>(0.07)</td>
</tr>
<tr>
<td>$\ln x_2$</td>
<td>1.23</td>
<td>(0.08)</td>
</tr>
<tr>
<td>$\ln x_1 \ln x_1$</td>
<td>-0.05</td>
<td>(0.001)</td>
</tr>
<tr>
<td>$\ln x_2 \ln x_2$</td>
<td>-0.04</td>
<td>(0.01)</td>
</tr>
<tr>
<td>$\ln x_1 \ln x_2$</td>
<td>-0.14</td>
<td>(0.02)</td>
</tr>
</tbody>
</table>

| Adj. $R^2$            | 0.998       |          |

Data for the regression can be obtained from Table 1. The dependent variable is $\ln S$.

Standard errors are reported in parentheses. All parameters are significant at the 1 percent level.
### Table 3: Example 2 Data Descriptive Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Inputs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_1$</td>
<td>103.30</td>
<td>26.92</td>
<td>35.56</td>
<td>170.88</td>
</tr>
<tr>
<td>$x_2$</td>
<td>100.13</td>
<td>24.85</td>
<td>30.76</td>
<td>154.86</td>
</tr>
<tr>
<td><strong>Outputs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y_1$</td>
<td>78.87</td>
<td>31.19</td>
<td>30.36</td>
<td>167.09</td>
</tr>
<tr>
<td>$y_2$</td>
<td>83.54</td>
<td>35.18</td>
<td>28.98</td>
<td>172.40</td>
</tr>
<tr>
<td><strong>Add. Variables</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_3$</td>
<td>101.72</td>
<td>27.22</td>
<td>45.22</td>
<td>164.99</td>
</tr>
<tr>
<td>$x_4$</td>
<td>101.95</td>
<td>24.63</td>
<td>47.44</td>
<td>154.86</td>
</tr>
<tr>
<td>$x_5$</td>
<td>101.53</td>
<td>27.94</td>
<td>39.98</td>
<td>175.39</td>
</tr>
<tr>
<td>Eff</td>
<td>0.85</td>
<td>0.09</td>
<td>0.62</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Descriptive statistics for example 2 data are calculated by the authors.
Table 4: Example 2 Regression Results

<table>
<thead>
<tr>
<th>Variable</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-4.85*</td>
<td>-5.15*</td>
<td>-5.44*</td>
<td>-5.42*</td>
</tr>
<tr>
<td></td>
<td>(0.23)</td>
<td>(0.30)</td>
<td>(0.36)</td>
<td>(0.40)</td>
</tr>
<tr>
<td>Ln x₁</td>
<td>0.35*</td>
<td>0.35*</td>
<td>0.35*</td>
<td>0.35*</td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>Ln x₂</td>
<td>0.62*</td>
<td>0.63*</td>
<td>0.63*</td>
<td>0.63*</td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>Ln x₃</td>
<td>----</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>Ln x₄</td>
<td>----</td>
<td>----</td>
<td>0.06</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>Ln x₅</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>-0.01</td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.04)</td>
</tr>
</tbody>
</table>

Data generation and estimation by the authors. The dependent variable is \( \ln S \).

Standard errors are reported in parentheses. * indicates significance at the 1 percent level. No other parameters were significant.
<table>
<thead>
<tr>
<th>Scenario</th>
<th>MAD</th>
<th>DEA</th>
<th>COLS</th>
<th>SDF</th>
<th>DEA</th>
<th>COLS</th>
<th>SDF</th>
<th>DEA</th>
<th>COLS</th>
<th>SDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.048</td>
<td>0.022</td>
<td>0.039</td>
<td>0.955</td>
<td>0.961</td>
<td>0.863</td>
<td>0.928</td>
<td>0.953</td>
<td>0.832</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.063</td>
<td>0.023</td>
<td>0.047</td>
<td>0.926</td>
<td>0.953</td>
<td>0.829</td>
<td>0.874</td>
<td>0.945</td>
<td>0.797</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.072</td>
<td>0.027</td>
<td>0.048</td>
<td>0.880</td>
<td>0.938</td>
<td>0.772</td>
<td>0.803</td>
<td>0.928</td>
<td>0.710</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.082</td>
<td>0.027</td>
<td>0.049</td>
<td>0.856</td>
<td>0.938</td>
<td>0.765</td>
<td>0.768</td>
<td>0.928</td>
<td>0.708</td>
<td></td>
</tr>
</tbody>
</table>

Randomly generated data for a two-output, two-input production process were used for the simulation. All calculations by authors. SDF is the approach popularized by Coelli and Perelman (2000). COLS is the multiple output corrected OLS approach developed in this paper.
Table 6: Random Effects Model Results

<table>
<thead>
<tr>
<th>Scenario</th>
<th>MAD</th>
<th>Correlation</th>
<th>Rank Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>One Output</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma_v = 0.15$</td>
<td>0.119</td>
<td>0.928</td>
<td>0.916</td>
</tr>
<tr>
<td></td>
<td>(0.06)</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>$\sigma_v = 0.20$</td>
<td>0.124</td>
<td>0.890</td>
<td>0.868</td>
</tr>
<tr>
<td></td>
<td>(0.07)</td>
<td>(0.01)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>$\sigma_v = 0.25$</td>
<td>0.121</td>
<td>0.844</td>
<td>0.816</td>
</tr>
<tr>
<td></td>
<td>(0.07)</td>
<td>(0.02)</td>
<td>(0.02)</td>
</tr>
<tr>
<td><strong>Two Outputs</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma_v = 0.15$</td>
<td>0.121</td>
<td>0.922</td>
<td>0.908</td>
</tr>
<tr>
<td></td>
<td>(0.07)</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>$\sigma_v = 0.20$</td>
<td>0.120</td>
<td>0.882</td>
<td>0.860</td>
</tr>
<tr>
<td></td>
<td>(0.07)</td>
<td>(0.01)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>$\sigma_v = 0.25$</td>
<td>0.113</td>
<td>0.836</td>
<td>0.807</td>
</tr>
<tr>
<td></td>
<td>(0.07)</td>
<td>(0.02)</td>
<td>(0.03)</td>
</tr>
</tbody>
</table>

Results reported are averages from 100 replications. Standard deviations are reported in parentheses. All calculations by authors.
### Table 7: Fixed Effects Model Results

<table>
<thead>
<tr>
<th>Scenario</th>
<th>MAD</th>
<th>Correlation</th>
<th>Rank Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>One Output</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma_v = 0.15$</td>
<td>0.071</td>
<td>0.926</td>
<td>0.916</td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>$\sigma_v = 0.20$</td>
<td>0.096</td>
<td>0.880</td>
<td>0.868</td>
</tr>
<tr>
<td></td>
<td>(0.05)</td>
<td>(0.01)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>$\sigma_v = 0.25$</td>
<td>0.121</td>
<td>0.829</td>
<td>0.816</td>
</tr>
<tr>
<td></td>
<td>(0.06)</td>
<td>(0.02)</td>
<td>(0.02)</td>
</tr>
<tr>
<td><strong>Two Outputs</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma_v = 0.15$</td>
<td>0.074</td>
<td>0.918</td>
<td>0.908</td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>$\sigma_v = 0.20$</td>
<td>0.100</td>
<td>0.872</td>
<td>0.860</td>
</tr>
<tr>
<td></td>
<td>(0.05)</td>
<td>(0.01)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>$\sigma_v = 0.25$</td>
<td>0.124</td>
<td>0.821</td>
<td>0.807</td>
</tr>
<tr>
<td></td>
<td>(0.06)</td>
<td>(0.02)</td>
<td>(0.03)</td>
</tr>
</tbody>
</table>

Results reported are averages from 100 replications. Standard deviations are reported in parentheses. All calculations by authors.
Appendix

Proof of Theorem 1:

The logic of the proof is similar to Proposition 5 by Banker (1993) which established consistency of the DEA estimator and Theorem 3.1 of Johnson and Kuosmanen (2009). By assumption i) isoquants are nested and by assumption ii) output sets can be analyzed for a given aggregate input level. Consider an arbitrary randomly drawn observation \((y_i, x_i)\). For any arbitrary input level \(x_i\), there is a positive probability \(p_i > 0\) of randomly drawing to the sample an observation \(k\) such that:

\[
f(x_k) = W, v_k = V^M.
\]

For this observation \(y_k = W + V^M\). Note that since the boundary of \(T\) is a globally concave function, it is not possible to achieve a higher output level than \(y_k\) by using input vector \(x_k\). Thus, if an observation \(k\) characterized by the equations above is randomly drawn, then \(y_k\) is a member of the set \(Isoq(x_i)\). Otherwise, if the observation \(k\) is not drawn to the sample, \(y_k\) is not a member of \(P_A\). Consistency requires that the probability of drawing unit \(k\) approach unity as the sample size approaches infinity.

The probability that unit \(k\) is not observed in a sequence of \(n\) independent random draws is equal to \((1-p_i)^n\). Asymptotically, this probability converges to zero:

\[
\lim_{n \to \infty} (1 - p_i)^n = 0.
\]

Thus, observation \(k\) is almost surely observed as the sample size approaches infinity. Hence

\[
\lim_{n \to \infty} Isoq(x_i) = Isoq(P_a) + V^M.
\]

As the argument was made for an arbitrary \(x_i\), the same argument can be made for any observation \(i = 1, \ldots, n\). This shows the true isoquant augmented by a noise component can be consistently estimated. The true isoquant can be recovered by subtracting \(V^M\).
References


