Outliers in data envelopment analysis

Taylor Boyd
Department of Economics and Finance, University of Dayton, Dayton, Ohio, USA

Grace Docken
Department of Chemical Engineering, University of Dayton, Dayton, Ohio, USA, and

John Ruggiero
Department of Economics and Finance, University of Dayton, Dayton, Ohio, USA

Abstract

Purpose – The purpose of this paper is to improve the estimation of the production frontier in cases where outliers exist. We focus on the case when outliers appear above the true frontier due to measurement error.

Design/methodology/approach – The authors use stochastic data envelopment analysis (SDEA) to allow observed points above the frontier. They supplement SDEA with assumptions on the efficiency and show that the true frontier in the presence of outliers can be derived.

Findings – This paper finds that the authors' maximum likelihood approach outperforms super-efficiency measures. Using simulations, this paper shows that SDEA is a useful model for outlier detection.

Originality/value – The model developed in this paper is original; the authors add distributional assumptions to derive the optimal quantile with SDEA to remove outliers. The authors believe that the value of the paper will lead to many citations because real-world data are often subject to outliers.

Keywords Performance, Outliers, Econometrics, Stochastic DEA, Efficiency estimation

Paper type Research paper

1. Introduction

Measurement of productive performance requires estimation of frontier production to serve as a benchmark for observed production. Frontier production is defined as the maximum output that can be produced with a given level of resources. Comparing the frontier production to the observed production provides a measure of decision-making unit (DMU) efficiency; inefficient units could increase output without an increase in inputs. There are many different approaches that have been used to estimate frontier production. A common

This article is part of a special issue Guest Edited by Rajiv D. Banker and Vincent Charles.

Journal of Centrum Cathedra: The Business and Economics Research Journal
Vol. 9 No. 2, 2016
pp. 168-183
Emerald Group Publishing Limited
1851-6599
DOI 10.1108/JCC-09-2016-0010

approach in the one-output setting is to assume a functional form of production consistent with beliefs of the underlying production process. If all deviations from the frontier (function) are due to statistical noise generated from a symmetric distribution centered at zero, then ordinary least squares (OLS) would be an appropriate technique subject to the usual econometric problems (e.g. simultaneity resulting from observed inputs being correlated with the unobserved inefficiency term). Alternatively, one could estimate this production function using nonparametric regression[1]. Collier *et al.* (2016) estimate a sports production function using alternative parametric and nonparametric estimation approaches. The nonparametric regressions remove the requirement of specification of the functional form but come at a cost of slower convergence.

If, however, the data-generating process is characterized instead by a one-sided error term that reflects only inefficiency, one could use OLS and adjust the intercept by the largest residual. This approach was suggested by Winsten (1957) and Richmond (1974). As discussed by Aigner *et al.* (1977) and proven by Greene (1980), the approach works because the estimated slope parameters are consistent and unbiased. Shifting the intercept leads to a consistent estimate of the intercept. Of course, the regression-based approach is parametric and is subject to model misspecification and biases if the true functional form of production is not correctly chosen. Additionally, the approach typically allows only one dependent variable, and hence, studies use an aggregate output (leading to potential biases), or requires additional information on prices to estimate the associated cost function[2].

Alternatively, one could use data envelopment analysis (DEA) developed by Charnes *et al.* (1978) and Banker *et al.* (1984). This deterministic model measures frontier production nonparametrically assuming general axioms of production. One major advantage of DEA is the ability to simultaneously allow multiple inputs and multiple outputs; efficiency is typically measured using a Farrell (1957) measure of equiproportional reduction of inputs (or expansion of outputs) to evaluate observed production relative to frontier production. The Charnes, Cooper and Rhodes model assumed constant returns to scale, leading to efficiency estimates that combine technical and scale efficiency. The Banker, Charnes and Cooper (BCC) model allows convexity and, hence, variable returns to scale.

Initially, DEA was criticized by econometricians who objected to the assumption that all deviations from the frontier were due to inefficiency. Instead, if data were perturbed not only by inefficiency but also by statistical noise, the estimated frontier would be biased upward and estimates of inefficiency would be composed not only of true inefficiency but also measurement error and other statistical noise. Of course, this problem exists not only with DEA but also with regression-based approaches that assume that deviations from the frontier are only one-sided and due to inefficiency.

The stochastic frontier model, developed by Aigner *et al.* (1977), extends OLS by allowing a composed error consisting of measurement error and noise. Assuming a production function, the parametric model requires additional assumptions on distributions for statistical noise and inefficiency[3]. Typically, noise is modeled with a normal distribution and the inefficiency is typically assumed to be either half-normal or exponential. Importantly, the initial models allowed only one output and a priori specification of the production function and error components. Assuming correct selection, the model provides unbiased and consistent estimates of all parameters including the intercept. It has been shown by Greene (1980) that the frontier can be obtained by using OLS and adjusting the intercept based on moments of the

distribution. The amount of shift depends on the skewness of the OLS residuals. In a cross-sectional setting, while the model does a good job identifying the production frontier, attempts to measure efficiency have failed. Jondrow *et al.* (1982) provided the first measure of firm-specific inefficiency by estimating the expected value of inefficiency given the observed composed error. Ondrich and Ruggiero (2001) prove that the resulting efficiency measure is simply a rescaling of the observed error. Importantly, the rank of efficiency using this estimator is unchanged from the ranking of the error. Furthermore, like DEA, the performance declines as the variance of the measurement error increases, *ceteris paribus*.

An alternative to DEA, stochastic data envelopment analysis (SDEA) represents an attempt to estimate DEA under the assumptions of the stochastic frontier model. Such a model would prove useful by relaxing the a priori selection of a production function[4]. Banker (1988) and Banker and Maindiratta (1992) provided alternative DEA formulations to estimate a DEA-type frontier through the middle of the data in the context of the stochastic frontier model. One of the models contained in Banker and Maindiratta (1992) assumes a composed error term (e.g. a normal distribution for statistical noise and a half-normal distribution for inefficiency like the stochastic frontier approach.) Instead of solving a linear program for each observation (like DEA), the model formulates one DEA-type model with constraints that satisfy convexity and monotonicity for the predicted output, leading to a piecewise linear production frontier. The conversion to one model is important because the distance to the frontier could be influenced by the error of all other DMUs. The objective function is the likelihood that the data came from the assumed distribution. Conceptually, the model works well but is limited by the number of observations because the number of constraints is roughly the sample size squared. Nonetheless, with modern computing, the model can handle typical-sized problems.

The Banker and Maindiratta (1992) model provides a maximum likelihood estimator using the Afriat conditions[5] to maintain the axioms of DEA. These important models have been recently revived with the names convex nonparametric least squares (CNLS) and stochastic nonparametric envelopment of data (StoNED) (Kuosmanen, 2008; Kuosmanen and Johnson, 2010 and Kuosmanen and Kortelainen, 2012). In the case of CNLS, the objective is to minimize the sum of squared residuals like OLS using the Afriat conditions. Of course, if the true error is a normal distribution with mean zero, the resulting CNLS model reduces to the Banker and Maindiratta (1992) maximum likelihood estimator, assuming the likelihood function of the error terms is normally distributed. Banker (1988) provided a nonparametric quantile regression that seeks to minimize the sum of absolute errors subject to the Afriat conditions, leading to a piecewise production function estimator that allowed observed points to appear above the frontier. If the error distribution is LaPlace (double exponential) centered at zero, then the median quantile using Banker's model would be maximum likelihood. The StoNED model combines the known intercept shift together with the Banker and Maindiratta (1992) model under an assumption of normally distributed errors.

Banker (1988) provides the foundation for an SDEA model where the user pre-specifies the quantile but does not provide a way to choose the optimal quantile. In practice, it is not known what the optimal quantile is. Banker *et al.* (2013) analyzed the sensitivity of the SDEA model in cases where inputs and/or outputs are perturbed. In particular, sufficient and necessary conditions are provided where perturbed data do not change the efficiency scores in the SDEA model. In this paper, we add an additional assumption on the distribution of the error term but allow observed infeasible outliers

drawn from different distributions. For example, if we make the common assumption that the true distribution of inefficiency is half-normal, we are able to choose the appropriate quantile based on maximum likelihood criteria. As a result, Banker's SDEA model can be used to detect and remove outliers.

Banker and Gifford (1988) and Andersen and Petersen (1993) developed a DEA model to remove overly influential points in the construction of the production frontier. The super-efficiency model was recommended by Banker and Gifford (1988) and later by Wilson (1995) as a useful approach for handling outliers to provide a better approximation of the production frontier. Banker and Chang (2006) provided an analysis using simulated data to show that the super-efficiency can improve efficiency estimation in the case where there are a few observations contaminated with noise[6]. While the approach can help identify influential outliers, the decision on how many units to exclude is arbitrary. And as the number of outliers increases, the effectiveness of the approach diminishes because the probability of having outliers in the same neighborhood increases.

The purpose of our paper is to develop an alternative DEA model to detect outliers. We begin with Banker's SDEA model and estimate numerous quantiles. We then assume a distribution for inefficiency and choose the quantile from the SDEA model consistent with the assumed distribution of inefficiency. In particular, we seek the quantile where the resulting errors maximize the likelihood that the points below the frontier come from the assumed distribution. This method will prove useful for empirical analyses where infeasible outliers appear above the frontier from a different distribution than the inefficiency.

The rest of the paper is organized as follows. In Section 2, we present the empirical production possibility set and show the estimated frontier using DEA. We also show the bias introduced in the frontier in the presence of outliers. In Section 3, we present Banker's SDEA model and discuss the distribution and likelihood function for the inefficiency component. We then provide a useful algorithm to estimate the frontier and inefficiency in the presence of outliers. Section 4 provides simulation analyses based on differing assumptions of error variances, per cent of outliers and sample size. Section 5 concludes with directions for future research.

2. Data envelopment analysis and outliers

We begin by defining the production possibility set assuming that there are no outliers. Assume that each of N DMUs produces one output y using a vector of M inputs $x = (x_1,\ldots, x_M)$. Output and input levels for DMU i are given by y_i and $x_{1i},\ldots, x_{Mi}$, respectively. The technology can be represented by the following empirical production possibility set:

$$T = \left\{(x,y) \,|\, \sum_{i=1}^{N} \lambda_i y_i \geq y, \; \sum_{i=1}^{N} \lambda_i x_{ji} \leq x_j, \text{ for } j = 1,\ldots,M, \; \sum_{i=1}^{N} \lambda_i = 1 \; \lambda_i \geq 0, \text{ for } i = 1,\ldots,N\right\}. \tag{1}$$

This is the standard production possibility set used in DEA consistent with the variable returns to scale the model of Banker *et al.* (1984).

Figure 1 illustrates the production frontier assuming one input and one output. Input data x_1 were generated U(1,10) with efficient production given by $y^* = 2x_1^{0.4}$. To obtain observed output allowing inefficiency, we generated $\ln u \sim |N(0,0.2)|$; observed output was calculated by multiplying efficient production by the efficiency index $e^{-\ln u}$. The BCC variable returns to scale the DEA model to measure output-oriented efficiency for DMU i is given by the following linear program:

$$\begin{aligned} &Max\ \theta_i \\ &s.t. \\ &\sum_{j=1}^{N} \lambda_j y_j \geq \theta_i y_i, \\ &\sum_{j=1}^{N} \lambda_j x_{mj} \leq x_{mi}, \text{ for } m = 1, \ldots, M, \\ &\sum_{j=1}^{N} \lambda_j = 1, \\ &\lambda_j \geq 0, \text{ for } j = 1, \ldots, N. \end{aligned} \tag{2}$$

The solution of linear program (Model 2) for each DMU provides a measure of output expansion possible given the technology in Model (1). Returning to our previous example, we solve the DEA model for each DMU and estimate the frontier production. The resulting frontier, shown in Figure 2, is a piecewise linear approximation to the true function that generated the data. The true frontier and the estimated DEA frontier are superimposed in Figure 2; with 100 observations, the nonparametric DEA estimator does a good job in identifying the true frontier. We note that the estimated frontier is biased downward due to sample size.

Using the same data, we generate outliers that are above the frontier. In particular, we invert the error term (in multiplicative form) for the first 10 points.

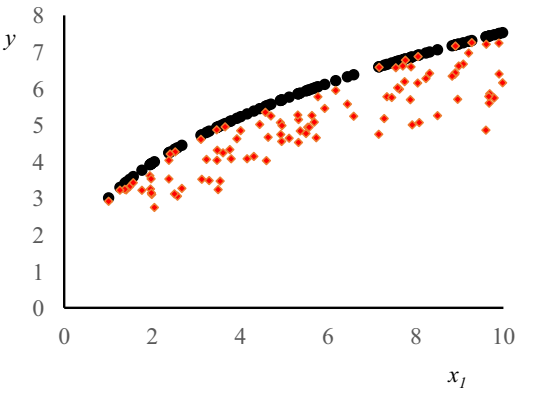

Figure 1.
Illustrative data

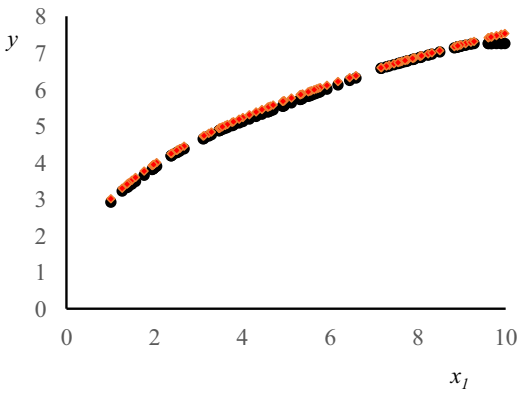

Figure 2.
DEA-estimated frontier versus true frontier

These data are illustrated in Figure 3. We note that two of the points are close to the frontier, but there are now several relatively large positive residuals. The effect on the estimated DEA frontier become obvious; outliers caused by positive errors lead to a biased frontier. Given that the frontier is biased upward, the associated DEA estimates of technical efficiency will be biased downward. DMUs will appear to be more inefficient than they truly are. We also note that the outliers will cause a distortion in the relative ranking of efficiency. In this specific example, DMUs with low input levels are not severely impacted by the presence of the outliers. However, as input levels increase, inefficiency becomes increasingly biased downward.

One way of detecting outliers is to measure super-efficiency, a procedure developed by Banker and Gifford (1988) to remove overly influential outliers. Andersen and Petersen (1993) applied the approach to rank efficient units. Wilson (1995) provided an alternative way to use the super-efficiency model to detect and remove outliers. The super-efficiency measure for DMU i is obtained with a modification to Model (2) by not allowing a given unit to serve in the benchmark for itself:

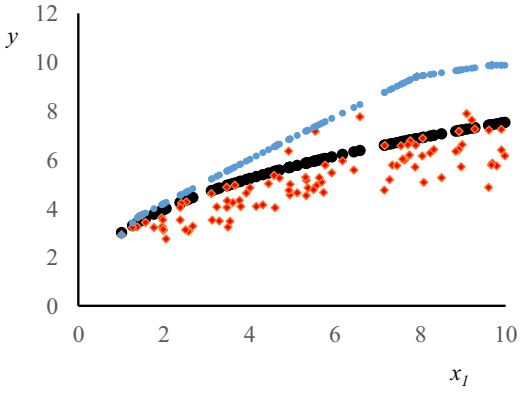

Figure 3.
Data with outliers

$$
\begin{aligned}
&Max\ \eta_i \\
&s.t. \\
&\sum_{j=1}^{N} \lambda_j y_j \geq \eta_i y_i, \\
&\sum_{j=1}^{N} \lambda_j x_{mj} \leq x_{mi}, \text{ for } m = 1, \ldots, M, \\
&\sum_{j=1}^{N} \lambda_j = 1, \\
&\lambda_i = 0, \\
&\lambda_j \geq 0, \text{ for } j = 1, \ldots, N.
\end{aligned} \tag{3}
$$

If $\theta_i^* = 1$ from Model (2) and $\eta_i^* < 1$ from Model (3), DMU i was identified as efficient because there were no other feasible convex combinations other than itself. Comparing Models (2) and (3), we have a way to evaluate the extent to which a unit's efficiency measure is influenced by allowing it to serve in the reference set. However, the influential DMU is still included in the reference set for all other DMUs. Wilson (1995) provides a modified model to identify the influence not only on itself but also on all other units that were evaluated relative to DMU i. We consider the following model[7]:

$$
\begin{aligned}
&Max\ \eta_i^{(k)} \\
&s.t. \\
&\sum_{j=1}^{N} \lambda_j y_j \geq \eta_i^{(k)} y_i, \\
&\sum_{j=1}^{N} \lambda_j x_{mj} \leq x_{mi}, \text{ for } m = 1, \ldots, M, \\
&\sum_{j=1}^{N} \lambda_j = 1, \\
&\lambda_k = 0, \\
&\lambda_j \geq 0, \text{ for } j = 1, \ldots, N.
\end{aligned} \tag{4}
$$

Here, DMU k is removed from the reference set and the output-oriented DEA model is solved. We note that Model (4) only needs to be solved for units that are identified as efficient in Model (2), as an inefficient unit cannot serve as a benchmark in Model (2) and, hence, has no effect on Model (4). We can then measure the influence that each unit has on the measurement of efficiency for the remaining DMUs by calculating the average difference in efficiency scores:

$$
\delta_k = \frac{1}{N-1} \sum_{i=1}^{N} (\eta_i - \eta_i^{(k)}). \tag{5}
$$

Large values of δ_k indicate that DMU k has a large average influence in the calculation of efficiency. Unfortunately, there are no guidelines on how Model (5) can be used to

remove outliers. Large values of δ_k are indicative of large influence, but this can arise if the unit is an infeasible outlier or one that is feasible but relatively more efficient than other units. As a consequence, using this method with an arbitrary decision rule might remove useful information from the sample.

We apply the super-efficiency model for outlier detection for the data shown in Figure 3. We define an arbitrary decision rule to delete observations if $\delta_k \geq 0.1$ and iterate through until we obtain a sample where $\delta_k < 0.1$ for all remaining DMUs. In this case, we were able to effectively remove eight DMUs after seven iterations. The resulting frontier is shown in Figure 4. With the structure of outlier from this example, the super-efficiency outlier approach works well.

We now consider an alternative data set. For 100 observations, we generate inputs uniformly on the interval (1,10) with efficient production given by $y^* = 3x_1^{0.4}$. For 80 per cent of the data, we generate inefficiency from $\ln u \sim |N(0,0.15)|$. For the other observations, outliers are drawn from $-\ln u \sim |N(0,0.1)|$. The negative sign is included to place the points above the frontier. Data for this example are shown in Figure 5.

We used the same decision rule used in the first example (i.e. delete observations if the parameter $\delta_k \geq 0.1$) and derived the frontier after removing outliers. Only two

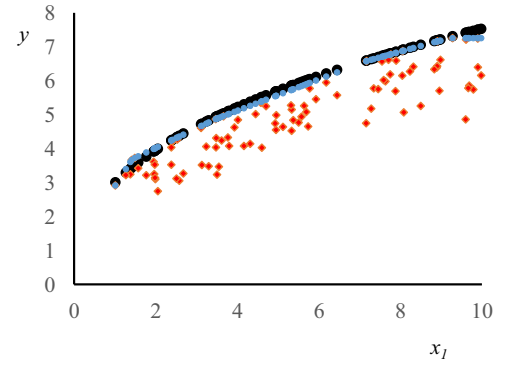

Figure 4.
Super-efficiency outlier detection

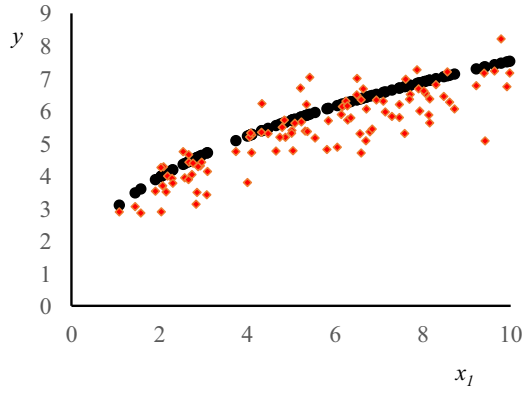

Figure 5.
Alternative data generation with outliers

DMUs were identified as outliers after only one iteration. The resulting frontier is shown in Figure 6 after removing the two outliers. In this case, the approach fails and the resulting frontier is biased with the inclusion of many outlier points. The failure is due to the proximity of outliers; removal of an outlier will not seriously affect super-efficiency scores because a neighboring outlier can serve as a benchmark. We note too that the decision rule for outlier detection is *ad hoc* and could in fact remove true frontier points that should be benchmarks.

3. Stochastic data envelopment analysis

We now turn to an alternative approach using the stochastic DEA model with additional information provided with an assumed distribution for inefficiency. Recognizing that deviations from the production frontier could be two-sided due to measurement, Banker (1988) formulates the DEA problem consisting of N linear programs into one linear program. Defining ε_{1i} and ε_{2i} as the positive and negative residual, respectively, for DMU i, the stochastic DEA model[8] seeks to minimize the sum of absolute residuals for a given quantile $0 < \tau < = 1$:

$$
\begin{aligned}
&Min\ \tau \sum_{j=1}^{N} \varepsilon_{1j} + (1 - \tau) \sum_{j=1}^{N} \varepsilon_{2j} \\
&s.t. \\
&\quad y_j = \alpha_j + \sum_{m=1}^{M} \beta_{mj} x_{mj} + \varepsilon_{1j} - \varepsilon_{2j} \quad \forall j, \\
&\quad \alpha_j + \sum_{m=1}^{M} \beta_{mj} x_{mj} \leq \alpha_k + \sum_{m=1}^{M} \beta_{mk} x_{mj} \quad \forall j, k, \\
&\quad \varepsilon_{1j}, \varepsilon_{2j} \geq 0 \quad \forall j.
\end{aligned}
\tag{6}
$$

The first set of constraints defines observed production as predicted frontier production, accounting for positive and negative residuals. Because the residuals are defined to be positive, the negative residual ε_{2i} for each DMU i is subtracted from frontier production. In the solution to Model (6), it is clear that both residual terms for each DMU cannot be

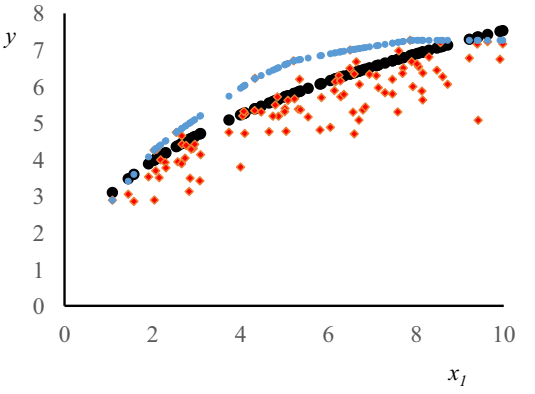

Figure 6.
Super-efficiency outlier detection

positive. The second set of constraints are the Afriat constraints that ensure convexity. Each DMU's predicted output evaluated using its own parameters is lower than the predicted output evaluated using all other DMUs' parameters. For the given τ, we obtain a piecewise linear production frontier where τ determines the percentile of the overall error distribution.

One limitation of the stochastic DEA model is the choice of τ. Without additional information, it is not clear what value is appropriate. For our purposes, we will provide the additional information with an assumption on the distribution of the inefficiency component. We will assume that the inefficiency (in natural log units) is distributed half-normally: $u \sim |N(0,\sigma)|$. The probability density function is given by:

$$f(u\,|\,\sigma) = \frac{\sqrt{2}}{\sigma\sqrt{\pi}} e^{\frac{-u^2}{2\sigma^2}}. \tag{7}$$

With N_1 observations, the log-likelihood function is:

$$L = \frac{N_1}{2}\ln(2 - \pi) - N_1 \ln \sigma - \frac{1}{2\sigma^2}\sum_{j=1}^{N_1} u^2. \tag{8}$$

Without any outliers, the DEA estimator provides a useful nonparametric estimator of the production frontier[9].

The problem becomes more complicated when infeasible points are observed above the frontier due perhaps to measurement error. We assume that $N_1 = \delta N$ with $\delta < 1$ of the data points are drawn from the half-normal distribution, while the remaining $N - N_1$ points are randomly drawn to be above the frontier. We do not require any assumptions on the distribution of the outlier points; our simulations in the next section consider a uniform distribution and a half-normal. Our goal is to find the optimal τ^* that maximizes the likelihood that the points below the frontier are drawn from a half-normal distribution. We propose solving the SDEA model for numerous quantiles and choosing the one that leads to maximum likelihood.

For illustrative purposes, we apply the approach to the outlier data shown in Figure 3. We solved all SDEA models with $\tau = 0.7$ to 0.995 in increments of 0.005. The resulting optimal $\tau^* = 0.845$. The resulting estimate of the production frontier is shown in Figure 7. We observe the approach has corrected the bias caused by the outliers and does a much better job in estimating the production frontier and is comparable to the case when outliers were not present in the data and to the super-efficiency approach discussed earlier. We next consider applying the model to Figure 5 data. Similar to the previous example, we solved all SDEA models with $\tau = 0.7$ to 0.995 in increments of 0.005. For this example, the optimal $\tau^* = 0.82$. (Figure 8).

We note that the likelihood approach produces a much better approximation to the frontier than the outlier approach. In the next section, we consider our approach using simulated data.

4. Simulations

For our simulations, we assume that the production frontier is given by $y^* = 3x_1^{0.4}$. We generate input $x_1 \sim U(1,10)$ for N observations. Next we choose the fraction δ of points

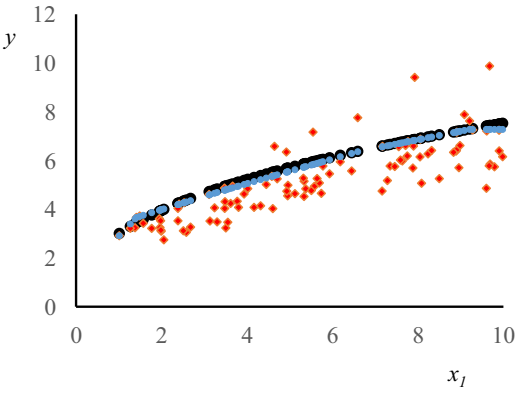

Figure 7.
Optimal SDEA frontier example 1

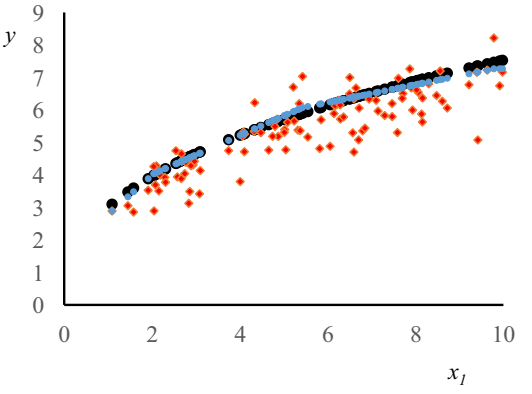

Figure 8.
Optimal SDEA frontier example 2

that do not contain outliers and generate efficiency for these points with $\ln \mu \sim |N(0,\sigma_\mu)|$. For the simulations, we consider three sample sizes: $N = 50$, 100 or 200. We also consider two different values for σ_μ, 0.1 or 0.2. The per cent of outliers is controlled by choosing two different levels of δ, 0.8 or 0.9. For the simulations, we randomly generate a uniform variable between 0 and 1 and determine if the observation is an outlier based on comparisons to the chosen δ.

For the outliers, two different distributions are selected. First, we consider $\ln \nu \sim -|N(0,\sigma_v)|$. Here, the points above the frontier, like the points below the frontier, are generated from a half-normal distribution. We vary σ_v with $\sigma_v = 0.1$ or 0.2. We also consider a uniform distribution with $\ln \nu \sim U(v_l,0)$ and v_l taking on two values: $v_l = -0.1$ or -0.15. Observed output is calculated as $y = e^{y^* - \ln \nu}$ ($y = e^{y^* - \ln u}$) for the outliers (non-outliers).

The full design has 48 simulations; for each choice of the outlier distribution, there are 24 simulations depending on three choices each for the sample size and two choices each for the per cent of outliers δ, and the standard deviations of the error components σ_μ and σ_v for the half-normal case and v_l for the uniform case.

For each simulation, we solved for all quantiles from 0.995 to 0.70 in increments of 0.005. In each case, we calculated the natural log of the ratio of predicted to observed output for those points that were below the frontier. The log likelihood function was evaluated, and the optimal τ^* was selected as the τ associated with the maximum log likelihood. Predicted values for each observation were then chosen from the values associated with τ^*. Unlike the super-efficiency model, the likelihood-based model does not eliminate any points.

As a benchmark, we also calculate the predicted value obtained from the DEA model for the non-outlier points. Of course, in real applications, you do not know which points are outliers; the results are only meant for comparison purposes. In Table I, we present the simulation results under the scenario that the outliers were generated from a half-normal distribution. Each line represents one of the 24 simulations. We sort the simulations in increasing order by the sample size, σ_μ, σ_v and δ. The mean squared error (MSE) between the true output level and the predicted output level is reported for the likelihood outlier approach and for DEA on the reduced sample. The true production output takes on values between 3 and approximately 7.5. For the average input value, output is approximately 5.93; an error of 0.03 is relatively small relative to the output values.

				Mean squared error	
N	σ_μ	σ_v	δ	Outlier approach	DEA[a]
50	0.1	0.1	0.8	0.028	0.025
			0.9	0.032	0.019
		0.2	0.8	0.006	0.008
			0.9	0.015	0.008
	0.2	0.1	0.8	0.030	0.017
			0.9	0.036	0.045
		0.2	0.8	0.053	0.039
			0.9	0.051	0.032
100	0.1	0.1	0.8	0.005	0.008
			0.9	0.023	0.002
		0.2	0.8	0.013	0.008
			0.9	0.003	0.003
	0.2	0.1	0.8	0.043	0.031
			0.9	0.015	0.019
		0.2	0.8	0.035	0.008
			0.9	0.032	0.008
200	0.1	0.1	0.8	0.008	0.005
			0.9	0.011	0.002
		0.2	0.8	0.014	0.004
			0.9	0.005	0.003
	0.2	0.1	0.8	0.036	0.008
			0.9	0.028	0.019
		0.2	0.8	0.044	0.011
			0.9	0.015	0.015

Note: [a]DEA was calculated using a subsample of data where outliers were excluded. The results are included as a benchmark

Table I.
Mean squared error between true and predicted, half-normal distribution

The results of the simulation are encouraging. While DEA on the reduced sample tends to achieve a lower MSE, the outlier approach provides comparable results. In 6 of the 24 cases, the likelihood approach provides a better fit than DEA, as indicated by a lower MSE. In the case of 200 observations, the absolute difference in MSE was less than 0.01 in all but one case. Focusing on the outlier approach, we do not observe a pattern comparing results across the percentage of outliers generated[10]. However, as the sample size increases, holding all other parameters fixed, we tend to see an improvement in the outlier approach. Furthermore, the outlier approach appears to provide similar results across standard deviation choices for both distributions, *ceteris paribus*.

The results for the uniform distribution are reported in Table II. We find that the approach provides quantitatively similar results as the half-normal simulations. In this case, the outlier approach achieves a smaller MSE in only 4 of the 24 simulations, but the differences tend to be smaller. As sample size increases, *ceteris paribus*, the outlier approach usually improves with a lower MSE. Overall, the likelihood approach is not sensitive to which distribution was chosen to generate the outliers.

				Mean squared error	
N	σ_μ	v_l	δ	Outlier approach	DEA[a]
50	0.1	0.1	0.8	0.012	0.012
			0.9	0.006	0.009
		0.2	0.8	0.012	0.018
			0.9	0.064	0.004
	0.2	0.1	0.8	0.038	0.054
			0.9	0.034	0.028
		0.2	0.8	0.058	0.025
			0.9	0.012	0.012
100	0.1	0.1	0.8	0.038	0.008
			0.9	0.011	0.007
		0.2	0.8	0.007	0.007
			0.9	0.009	0.006
	0.2	0.1	0.8	0.034	0.007
			0.9	0.088	0.014
		0.2	0.8	0.004	0.008
			0.9	0.028	0.033
200	0.1	0.1	0.8	0.033	0.003
			0.9	0.006	0.004
		0.2	0.8	0.006	0.004
			0.9	0.006	0.003
	0.2	0.1	0.8	0.009	0.010
			0.9	0.010	0.004
		0.2	0.8	0.017	0.009
			0.9	0.013	0.005

Table II.
Mean squared error between true and predicted, uniform distribution

Note: [a]DEA was calculated using a subsample of data where outliers were excluded. The results are included as a benchmark

5. Conclusions

In this paper, we presented the outlier detection model using the super-efficiency. The approach works well when there tends to be few outliers or if the structure of the outliers allows super-efficiency measures to determine influence. This is not always the case. In particular, when there are a lot of outliers, there is an increased probability that the outliers have neighbors that are also outliers. In this case, removal of one outlier has little effect on estimation because the neighbor outlier serves as a proxy benchmark.

We developed an alternative method based on the stochastic DEA model of Banker (1988). The model provides a DEA-type quantile regression that chooses the most likely DEA frontier for a given quantile. We supplement this model by assuming a distribution for inefficiency and find the quantile that maximizes the likelihood that the points below the frontier were generated from the assumed distribution. We provided an example where the likelihood approach performed extremely well, while the super-efficiency method did not. Both methods require an iterative approach. The super-efficiency approach requires an arbitrary decision rule for identifying influential outliers for removal and subsequent re-estimation of the frontier. The likelihood approach iterates over all quantiles and chooses the optimal quantile based on an assumed probability distribution for inefficiency.

We conducted a simulation analysis of the likelihood approach and allowed the number of observations, the standard deviation of inefficiency, the percentage of outliers and the distribution for the outliers to vary. The results indicated that the likelihood approach performs well, achieving comparable results to DEA models across specifications. Future research could test the sensitivity to incorrect specification of the inefficiency distribution and robustness in a more complete Monte Carlo analysis. In addition, the likelihood approach could be extended to estimate a true stochastic frontier where measurement error affects points above and below the frontier.

Notes

1. See, for example, Hall and Huang (2001), who use kernel regression while imposing monotonicity, and Du *et al.* (2013), who impose monotonicity and convexity. Henderson and Parmeter (2009) provide useful references and discussion.
2. One could also use a directional distance function and estimate the multiple-output, multiple-input equation.
3. Parmeter and Kumbhakar (2014) provide an excellent discussion of stochastic frontier analysis.
4. While Banker (1988) appears to be the earliest to coin the phrase "Stochastic DEA", other estimators use the phrase. For example, Simar and Zelenyuk (2011) also develop a stochastic DEA estimator to account for noise.
5. See Afriat (1967, 1972).
6. Banker and Chang (2006) also show that the super-efficiency measure is not useful for ranking DMUs by efficiency.
7. The variable returns to scale (VRS) BCC model could lead to infeasible points. One could use the modification proposed to Banker and Chang (2006). We follow Wilson (1995) and use information from feasible solutions.

8. The extension to the CNLS version of (6) is straightforward. Instead of minimizing the sum of absolute errors, we could instead minimize the sum of squared residuals. In the later sections, we solved both versions but only report the SDEA results. We find the differences to be insignificant and the model extension obvious.
9. Banker (1993) provided the statistical foundation for DEA and provides a useful discussion for our approach. Alternative distributions for inefficiency are available but are not considered in this paper. Future research could extend our method to alternative distributions consistent with Banker (1993).
10. Of course, a full Monte Carlo study should be used to determine the robustness across data-generating processes. We leave that for future work.

References

Afriat, S. (1967), "The construction of a utility function from expenditure data", *International Economic Review*, Vol. 8 No. 1, pp. 67-77.

Afriat, S. (1972), "Efficiency estimation of production functions", *International Economic Review*, Vol. 13 No. 3, pp. 568-598.

Aigner, D., Lovell, C.A.K. and Schmidt, P. (1977), "Formulation and estimation of stochastic frontier production function models", *Journal of Econometrics*, Vol. 6 No. 1, pp. 21-37.

Andersen, P. and Petersen, N.C. (1993), "A procedure for ranking efficient units in data envelopment analysis", *Management Science*, Vol. 39, pp. 1261-1264.

Banker, R. (1988), "Stochastic data envelopment analysis", Working Paper, Carnegie Mellon University, Pittsburgh, pp. 1-29.

Banker, R. (1993), "Maximum-likelihood, consistency and data envelopment analysis – a statistical foundation", *Management Science*, Vol. 39 No. 10, pp. 1265-1273.

Banker, R. and Chang, H. (2006), "The super-efficiency procedure for outlier identification, not for ranking efficient units", *European Journal of Operational Research*, Vol. 175 No. 2, pp. 1311-1320.

Banker, R. and Gifford, J. (1988), "A relative efficiency model for the evaluation of public health nurse productivity", Working Paper, Carnegie Mellon University, Pittsburgh.

Banker, R. and Maindiratta, A. (1992), "Maximum likelihood estimation of monotone and concave production frontiers", *Journal of Productivity Analysis*, Vol. 3 No. 2, pp. 401-415.

Banker, R., Charnes, A. and Cooper, W.W. (1984), "Some models for estimating technical and scale inefficiencies in data envelopment analysis", *Management Science*, Vol. 30 No. 9, pp. 1078-1092.

Banker, R., Kotarac, A. and Neralic, L. (2013), "Sensitivity and stability in stochastic data envelopment analysis", *Journal of the Operational Research Society*, Vol. 66 No. 1, pp. 1-14.

Charnes, A., Cooper, W. and Rhodes, E. (1978), "Measuring the inefficiency of decision making units", *European Journal of Operational Research*, Vol. 2, pp. 247-268.

Collier, T., Marquardt, K. and Ruggiero, J. (2016), "Nonparametric estimation of production functions", *Data Envelopment Analysis Journal*, Vol. 2, pp. 35-52.

Du, P., Parmeter, C. and Racine, J. (2013), "Nonparametric kernel regression with multiple predictors and multiple shape constraints", *Statistica Sinica*, Vol. 23 No. 3, pp. 1347-1371.

Farrell, M.J. (1957), "The measurement of productive efficiency", *Journal of the Royal Statistical Society: Series A. General*, Vol. 120 No. 3, pp. 253-281.

Greene, W. (1980), "Maximum likelihood estimation of econometric frontier productions", *Journal of Econometrics*, Vol. 13, pp. 27-56.

Hall, P. and Huang, H. (2001), "Nonparametric kernel regression subject to monotonicity constraints", *Annals of Statistics*, Vol. 29 No. 3, pp. 624-647.

Henderson, D. and Parmeter, C. (2009), "Imposing economic constraints in nonparametric regression: survey, implementation and extension", in Li, Q. and Racine, J. (Eds), *Advances in Econometrics: Nonparametric Econometric Methods*, Emerald Group Publishing, Bingley, Vol. 25, pp. 479-508.

Jondrow, J., Lovell, C.A.K., Materov, I.S. and Schmidt, P. (1982), "On the estimation of technical inefficiency in the stochastic frontier production function model", *Journal of Econometrics*, Vol. 19 Nos 2/3, pp. 233-238.

Kuosmanen, T. (2008), "Representation theorem for convex nonparametric least squares", *Econometrics Journal*, Vol. 11 No. 2, pp. 308-325.

Kuosmanen, T. and Johnson, A. (2010), "Data envelopment analysis as nonparametric least square regression", *Operations Research*, Vol. 58 No. 1, pp. 149-160.

Kuosmanen, T. and Kortelainen, M. (2012), "Stochastic non-smooth envelopment of data: semi-parametric frontier estimation subject to shape constraints", *Journal of Productivity Analysis*, Vol. 38 No. 1, pp. 11-28.

Ondrich, J. and Ruggiero, J. (2001), "Efficiency measurement in the stochastic frontier model", *European Journal of Operational Research*, Vol. 129 No. 2, pp. 435-442.

Parmeter, C. and Kumbhakar, S. (2014), "Efficiency analysis: a primer on recent advance", *Foundations and Trends in Econometrics*, Vol. 7 Nos 3/4, pp. 191-385.

Richmond, J. (1974), "Estimating the efficiency of production", *International Economic Review*, Vol. 15 No. 2, pp. 515-521.

Simar, L. and Zelenyuk, V. (2011), "Stochastic FDH/DEA estimators for frontier analysis", *Journal of Productivity Analysis*, Vol. 36 No. 1, pp. 1-20.

Wilson, P. (1995), "Detecting influential observations in data envelopment analysis", *Journal of Productivity Analysis*, Vol. 6 No. 1, pp. 27-45.

Winsten, C.B. (1957), "Discussion on Mr Farrell's paper", *Journal of the Royal Statistical Society Series A-Statistics in Society*, Vol. 120 No. 3, pp. 282-284.

Further reading

Parmeter, C. and Racine, J. (2012), "Smooth constrained frontier analysis", in Chen, X. and Swanson, N. (Eds), *A Festschrift in Honour of Halbert L. White Jr,* Springer, Berlin.

Corresponding author

John Ruggiero can be contacted at: jruggiero1@udayton.edu