
ROBUST EFFICIENCY MEASUREMENT

Dealing with outliers in Data Envelopment Analysis

Timo Kuosmanen

Helsinki School of Economics and Business Administration
Department of Economics and Management Science
P.O. Box 1210, FIN-00101 Helsinki, **FINLAND**

Thierry Post

Erasmus University Rotterdam
Department of Finance
Burg. Oudlaan 50
3062 PA Rotterdam, **THE NETHERLANDS**

March 1999

**Rotterdam Institute for Business Economic Studies (RIBES)
Report 9911**

All rights reserved. This study may not be reproduced in whole or in part without the authors' permission

Abstract

In this paper we propose a DEA model for estimating technical efficiencies if the data set contaminated with outliers. Excluding outliers restores the original analysis, but unfortunately the number and the identity of the outliers is typically unknown to the analyst. We propose to exclude the k most influential DMUs from the data set, where k is determined using empirical specification tests. We show that this procedure provides statistically consistent efficiency estimates in the case of outliers. In addition, Monte Carlo simulations suggest that the proposed approach can substantially improve the estimation in relatively small samples as well.

Keywords: Data Envelopment Analysis, Outliers, Robust Estimation

1. Introduction

Data Envelopment Analysis (DEA), outlined by Farrell (1957) and operationalized by Charnes, Cooper and Rhodes (1978), is a non-parametric method for estimating production frontiers and evaluating the relative efficiency of Decision Making Units (DMUs). The advantage of DEA over parametric stochastic frontier methods (Aigner et al. 1977, and Meeusen and van der Broeck 1977) has been its flexibility in multi-inputs multi-output environment, and robustness with respect to the specification of the functional relationships between inputs and outputs. However, since DEA relies on identifying best practice reference units, it can be extremely sensitive to outliers in the data set. In DEA framework, outliers are understood as observations lying outside the true production possibilities' set due to e.g. data error, heterogeneity of the DMUs, or erroneous production assumptions.

Stochastic DEA models that explicitly account for stochastic disturbances are a potential solution to this problem. In the DEA literature, a number of such models have been proposed (e.g. Gong and Sun 1995, Land et al. 1994; Olesen and Petersen 1995; Post 1997; Cooper et al. 1998, and Gstach 1998.). However, whether these techniques can deal with outliers remains as an open question. The stochastic models typically require the specification of a particular statistical distribution, and their robustness with respect to specification error has not been documented yet.

Another approach is to use preprocessing outlier detection routines, and to apply deterministic DEA models to the preprocessed data. Unfortunately, as pointed out by Wilson (1995), standard methods for detecting outliers in regression models can not be adopted to DEA, because DEA essentially has a deterministic structure. In addition, the literature on outlier detection in DEA is very sparse.

Wilson (1995) proposed one of the sparse outlier detection routines specially tailored to DEA. That procedure relies on assessing the impact of excluding observations from the data set. If the exclusion of a particular DMU has a large impact on the efficiency scores of the remaining DMUs, the input-output vector of that DMU is supported by few additional observations. Consequently, the observation is a potential outlier and it is assigned a high priority for follow-up inspection. A careful follow-up inspection of the data could reveal whether the observation has to be adjusted, omitted or can be included. This approach can very useful if data checking is costly and resources are scarce. However, the model relies on the ability of the analyst to identify the outliers from the set of prioritized observations. In addition, the modified model excludes the evaluated unit only. This approach may fail if the data set contains multiple

outliers. For example, the omission of an outlier can have little impact if one or more other outliers mask it.

Alternatively, outliers could be detected using techniques that measure the robustness of DEA efficient classifications. A number of such techniques have been proposed, e.g. Charnes et al. (1992), and Zhu (1996). Units with a very robust efficient classification are potential outliers, because the data set contains little additional empirical evidence to support the attainability of their input-output combination. However, like the Wilson procedure, the effectiveness of this approach depends on the success of the follow-up inspection. In addition, the models for assessing robustness essentially consider data variations for the evaluated unit only. Like the Wilson model, this approach may fail if the data set contains multiple outliers. For example, the classification of an outlier can be non-robust if other outliers mask it.

In this paper, we propose an alternative model for estimating efficiency if the data set is contaminated with outliers, the Robust Efficiency Model (REM). REM relies on excluding the set of most influential DMUs from the data set. We will demonstrate that this operation suffices to correct for the influence of outliers, without having to identify the outliers prior to the analysis. The proper number of exclusions can be selected using an empirical test procedure. Monte Carlo simulations suggest that REM can substantially improve the estimation relative to the standard DEA, and in addition that REM can come close to the 'ideal' solution of identifying and excluding all outliers.

The rest of the paper is organized as follows. Section 2 introduces the standard DEA model and discusses its sensitivity to outliers. Section 3 presents the Robust Efficiency Model. Section 4 uses Monte Carlo simulations to evaluate the finite sample performance of REM. Finally, section 5 offers concluding remarks and suggestions for future research.

2. The BCC model

DEA evaluates the efficiency of decision-making units relative to the production possibilities, and moreover identifies reference units that can help to find out causes and remedies for inefficiencies. Theoretically, the production possibilities can be represented by the production set:

$$(2.1) \quad T = \{(x, y) \in \mathfrak{R}_+^{m+s} \mid \text{input } x \text{ can produce output } y\}.$$

DEA models differ with respect to the assumptions imposed on the production possibilities, the measure used for evaluating efficiency, and the assumptions imposed on the statistical distribution of the observations in the data set. This study deals exclusively with the distribution assumptions, and restricts the production assumptions to the assumptions of the standard Banker, Charnes and Cooper (BCC, 1984) model. In addition, we restrict the efficiency measure to the Farrell (1957) input efficiency measure¹. Focussing on the evaluation of the j -th DMU, that measure is defined:

¹ However, the analysis applies directly to alternative production assumptions and efficiency measures.

$$(2.2) \quad \mathbf{q}_j = \min_{\mathbf{q}} \{ \mathbf{q} \mid (\mathbf{q}x_j, y_j) \in T \} \quad (x_j, y_j) \in T.$$

Unfortunately, the production set typically is unknown and has to be approximated using the observations in the data set. In this study, $X = (x_1 \dots x_n)^T$, with $x_j = (x_{1j} \dots x_{mj})$, and $Y = (y_1 \dots y_n)^T$, with $(y_{1j} \dots y_{sj})$, represent the data set, and $D = \{1, \dots, n\}$. In general, the production set is approximated as the smallest subset in input-output space that is consistent with the production and distribution assumptions imposed. The standard BCC model assumes that the production set satisfies free disposability and convexity, and in addition that the observations are contained within the production set. If these assumption hold, virtual input-output combinations created as convex combinations of observations are technically possible. Based on this insight, the BCC model uses the following empirical production set as an approximation for the true production set:

$$(2.3) \quad \hat{T}_{BCC} = \{(x, y) \mid x \leq \mathbf{I}^T X; y \geq \mathbf{I}^T Y; \mathbf{I}^T e = 1; \mathbf{I}_j \geq 0\}.$$

Measuring efficiency relative to this set gives the following efficiency estimator:

$$(2.4) \quad \hat{\mathbf{q}}_{BCC,j} = \min_{\mathbf{q}} \{ \mathbf{q} \mid (\mathbf{q}x_j, y_j) \in \hat{T}_{BCC} \} \\ = \min_{\mathbf{I}, \mathbf{q}} \{ \mathbf{q} \mid \mathbf{I}^T X \leq \mathbf{q}x_j; \mathbf{I}^T Y \geq y_j; \mathbf{I}^T e = 1; \mathbf{I} \geq 0 \}.$$

If the BCC assumptions hold, the approximating set is contained within the true production set, i.e. $\hat{T}_{BCC} \subseteq T$. Consequently, estimated efficiency is biased above true efficiency, i.e.

$$(2.5) \quad \hat{\mathbf{q}}_{BCC,j} \geq \mathbf{q}_j.$$

Therefore, $\hat{\mathbf{q}}_{BCC,j} = 1$ is a necessary condition for efficiency. In addition, introducing additional DMUs in the data set generally reduces bias. In fact, the estimator is statistically consistent (i.e. asymptotically unbiased and with a vanishing variance), if the observations are considered as identically and independently distributed random variables with a strictly positive density over the entire production set consistent (Banker 1993, Korostlev et al. 1995a, 1995b), i.e.:

$$(2.6) \quad P(x_j \leq x, y_j \geq y) = F(x, y) > 0 \quad \forall (x, y) \in T; j \in D.$$

These properties do not require the specification of a functional form for the production relationships, which gives DEA a comparative advantage relative to parametric stochastic frontier models (Aigner et al. 1977, and Meeusen and van der Broeck 1977), that typically do require such a specification.

However, as discussed above, the standard model assumes that all observations are contained within the production set. Outliers caused by e.g. heterogeneity of DMUs or erroneous production assumptions can violate this assumption. Such outliers can reduce the goodness of $\hat{\mathbf{q}}_{BCC,j}$ as an estimator for efficiency. More specifically, if the data set contains a

reference unit that produces more output than the evaluated unit and consumes less than the efficient amount of input for the evaluated unit, the estimator falls below true efficiency, i.e.:

$$(2.6) \quad \exists \mathbf{l} : \mathbf{l}^T X < \mathbf{q}_j x_j; \mathbf{l}^T Y \geq y_j; \mathbf{l}^T \mathbf{e} = 1; \mathbf{l} \geq 0 \Rightarrow \hat{\mathbf{q}}_{BCC,j} < \mathbf{q}_j.$$

If the estimator can fall below true efficiency, $\hat{\mathbf{q}}_{BCC,j} = 1$ is not a necessary condition for efficiency. In addition, increasing the number of observations can not reduce negative errors, and in fact introduces the risk of introducing additional outliers. Therefore $\hat{\mathbf{q}}_{BCC,j}$ generally is not statistically consistent in case of outliers.

3. The Robust Efficiency Model

If all observations are contained within the production set, condition (2.6) can not be satisfied. However, in contrast to the standard model, we assume here that the data set does contain observations outside the production set. Such outliers can be caused by e.g. heterogeneity of the DMUs, or erroneous production assumptions. We will denote the set of outliers by $Q = \{i \in D \mid (x_i, y_i) \notin T\}$.

Obviously, eliminating the outliers from the data set restores the original analysis. The remaining observations are contained within the production set, and hence the empirical production set is contained within the true production set. Unfortunately, the identity of the outliers is typically unknown, and, as discussed in the introduction, there is no unambiguous technique for detecting outliers in DEA.

However, we will demonstrate below that excluding the most influential DMUs from the data set suffices to correct for the influence of outliers, without having to identify the outliers prior to the analysis. In addition, we will demonstrate that a follow-up inspection of the excluded DMUs can detect outliers, which can further improve the estimation.

We propose to evaluate each DMU relative to a modification of the empirical production set \hat{T}_{BCC} , constructed by excluding a set of DMUs, say $K \subset D$, from the data set. The evaluated unit cannot be excluded itself, because that can give infeasible solutions. In addition, for the units in the production set, i.e. $j \in D \setminus Q$, including the evaluated unit effectively uses prior information to improve the estimation. We ignore the estimation of efficiency for the outliers, because efficiency is not well defined for units outside the production set. The reason is that $\{\mathbf{q} \mid (\mathbf{q}x_j, y_j) \in T\} \mid j \in Q$ can be empty, and in addition, if it is non-empty, $\mathbf{q}_j \mid j \in Q$ is a mixture of efficiency and productivity differences.

More specifically, using X^K to denote the matrix of inputs, the j^{th} ($j \in K$) rows omitted, and Y^K for the matrix of outputs, the j^{th} ($j \in K$) rows omitted, the evaluated DMU is evaluated against the following modified set:

$$(3.1) \quad \hat{T}_{BCC}^K = \{(x, y) \mid x \leq \mathbf{l}^T X^K; y \geq \mathbf{l}^T Y^K; \mathbf{l}^T \mathbf{e} = 1; \mathbf{l}_j \geq 0\} \quad K \subset D \mid j.$$

To exclude the most influential observations, we select K such that efficiency is maximized, giving the following Robust Efficiency Model (REM):

$$(3.2) \quad \hat{\mathbf{q}}_{REM,k,j} = \max_{K \subset D \setminus j} \min_{\mathbf{q}} \left\{ \mathbf{q} \mid (\mathbf{q}x_j, y_j) \in \hat{T}_{BCC}^K; \text{card}(K) \leq k \right\} \\ = \max_{K \subset D \setminus j} \min_{\mathbf{1}, \mathbf{q}} \left\{ \mathbf{q} \mid \mathbf{I}^T X^K \leq \mathbf{q}x_j; \mathbf{I}^T Y^K \geq y_j; \mathbf{I}^T \mathbf{e} = 1; \mathbf{I} \geq 0; \text{card}(K) \leq k \right\}.$$

Below, we will discuss the selection of the proper maximum number of exclusions, say $k = \text{card}(K)$. If that number is finite and greater or equal than the number of outliers, say $q = \text{card}(Q)$, (3.2) gives statistically consistent estimates for the DMUs in the production set, i.e.:

$$(3.3) \quad \lim_{n \rightarrow \infty} P\left(\hat{\mathbf{q}}_{REM,k,j} - \mathbf{q}_j \geq \mathbf{e}\right) = 0 \quad \forall \mathbf{e} > 0, \infty > k \geq q, j \in D \setminus Q.$$

Proof If $k \geq q$, eliminating the outliers only, i.e. $K = Q$, is a possible solution. Since $j \in D \setminus Q$, the associated efficiency estimator is well defined and corresponds to

$\min_{\mathbf{q}} \left\{ \mathbf{q} \mid (\mathbf{q}x_j, y_j) \in \hat{T}_{BCC}^Q \right\}$. Since $\hat{T}_{BCC}^Q \subseteq T$, this estimator is biased above true efficiency, i.e.

$P\left(\min_{\mathbf{q}} \left\{ \mathbf{q} \mid (\mathbf{q}x_j, y_j) \in \hat{T}_{BCC}^Q \right\} - \mathbf{q}_j \leq \mathbf{e}\right) = 0 \quad \forall \mathbf{e} > 0, k \geq q, j \in D \setminus Q$. In addition, since the optimal

K is selected by maximization, we have $\hat{\mathbf{q}}_{REM,k,j} \geq \min_{\mathbf{q}} \left\{ \mathbf{q} \mid (\mathbf{q}x_j, y_j) \in \hat{T}_{BCC}^Q \right\}$. Hence, $\hat{\mathbf{q}}_{REM,k,j}$ is biased above true efficiency as well, i.e.:

$$(I) \quad P\left(\hat{\mathbf{q}}_{REM,k,j} - \mathbf{q}_j \leq \mathbf{e}\right) = 0 \quad \forall \mathbf{e} > 0, k \geq q, j \in D \setminus Q.$$

In addition, if the data set contains more than k DMUs that consume at most $\mathbf{q}_j x_j$ to produce at least y_j , the modified production set necessarily contains at least one such DMU, and

hence $\hat{\mathbf{q}}_{REM,k,j}$ falls below true efficiency, i.e.:

$$(II) \quad \exists G \subseteq D : x_i \leq \mathbf{q}_j x_j \wedge y_i \geq y_j \quad \forall i \in G; \text{card}(G) > k \\ \Rightarrow \exists (x, y) \in \hat{T}_{BCC}^K : x \leq \mathbf{q}_j x_j \wedge y \geq y_j \Rightarrow \hat{\mathbf{q}}_{REM,k,j} \leq \mathbf{q}_j.$$

Since $j \in D \setminus Q$, we have $(\mathbf{q}_j x_j, y_j) \in T$. Therefore, using (2.6), the probability of observing

more than k of such DMUs equals $P\left(\exists G \subseteq D : x_i \leq \mathbf{q}_j x_j \wedge y_i \geq y_j \quad \forall i \in G; \text{card}(G) > k\right)$

$$= 1 - \sum_{i=1}^k \binom{n}{i} F(\mathbf{q}_j x_j, y_j)^i (1 - F(\mathbf{q}_j x_j, y_j))^{n-i}. \text{ Since } F(\mathbf{q}_j x_j, y_j) > 0 \text{ (2.6) and } k < \infty, \text{ this}$$

probability converges to unity as the sample increases, i.e.:

$$\lim_{n \rightarrow \infty} P\left(\exists G \subseteq D : x_i \leq \mathbf{q}_j x_j \wedge y_i \geq y_j \quad \forall i \in G; \text{card}(G) > k\right) =$$

$$\lim_{n \rightarrow \infty} \left[1 - \sum_{i=1}^k \binom{n}{i} F(\mathbf{q}_j x_j, y_j)^i (1 - F(\mathbf{q}_j x_j, y_j))^{n-i} \right] = 1. \text{ Combining this with (II), we find that the}$$

estimator asymptotically is biased below true efficiency, i.e.:

$$(III) \quad \lim_{n \rightarrow \infty} P\left(\hat{\mathbf{q}}_{REM,k,j} - \mathbf{q}_j \geq \mathbf{e}\right) = 0 \quad \forall \mathbf{e} > 0, k < \infty, j \in D \setminus Q.$$

Combining (I) and (III) gives consistency:

$$\lim_{n \rightarrow \infty} \left[P \left(\left| \hat{\mathbf{q}}_{REM,k,j} - \mathbf{q}_j \right| \geq \mathbf{e} \right) \right] = 0 \quad \forall \mathbf{e} > 0, q \leq k < \infty, j \in D \setminus Q.$$

Note that this property does not require the actual identification of the outliers. However, there are two complications.

First, consistency holds for DMUs in the production set only. Outliers can exclude all other DMUs outside the production set and hence receive a unity efficiency estimate, i.e.:

$$(3.4) \quad \hat{\mathbf{q}}_{REM,k,j} = 1 \quad \infty > k \geq q, j \in Q.$$

It is not clear how to interpret this estimate, because, as discussed above, efficiency relative to the production set is ill defined and meaningless for units outside the production set.

Second, consistency holds for all $k \geq q$. However, since $\hat{\mathbf{q}}_{REM,k,j}$ is nested in $\hat{\mathbf{q}}_{REM,p,j}$, $p > k$, we find $\hat{\mathbf{q}}_{REM,p,j} \geq \hat{\mathbf{q}}_{REM,k,j} \geq \mathbf{q}_j$ if $p > k \geq q$. Consequently, a large number of exclusions costs higher finite sample error, i.e.:

$$(3.5) \quad \left| \hat{\mathbf{q}}_{REM,k,j} - \mathbf{q}_j \right| \leq \left| \hat{\mathbf{q}}_{REM,p,j} - \mathbf{q}_j \right| \quad q \leq k < p.$$

Therefore, it is desirable to select k as small as possible (obviously, provided $k \geq q$).

In fact the ideal solution in terms of finite sample error is the omission of the outliers only, that is $K = Q$:

$$(3.6) \quad \left| \min_q \left\{ \mathbf{q} \left(\mathbf{q}x_j, y_j \right) \in \hat{T}_{BCC}^Q \right\} - \mathbf{q}_j \right| \leq \left| \hat{\mathbf{q}}_{REM,k,j} - \mathbf{q}_j \right| \quad k \geq q.$$

We propose to select the appropriate number of exclusions by using an empirical specification test. It follows from the above discussion, that both $\hat{\mathbf{q}}_{REM,k,j}$ and $\hat{\mathbf{q}}_{REM,p,j}$ are asymptotically distributed as true inefficiency if $p > k \geq q$. However, if $p \geq q > k$, we find

that $\hat{\mathbf{q}}_{REM,k,j}$ asymptotically can exceed $\hat{\mathbf{q}}_{REM,p,j}$. Consequently, in large samples, testing whether the two models give significantly different efficiency estimates can test the null hypothesis $H_0 : k \geq q$ against the alternative hypothesis $H_a : k < q$, under the maintained assumption that $p \geq q$.

The appropriate test statistic depends on the distribution of true efficiency. For example, Banker (1993) demonstrated that if the inefficiencies (i.c. $(1 - \mathbf{q}_j)$) are considered as identically and independently distributed random variables from a half-normal distribution, the ratio of the sum of squared inefficiencies over all DMUs for two nested models asymptotically follows an F-distribution, provided both models are consistent. The distribution assumptions are a subject of considerable debate, and alternative test statistics have been proposed (e.g. Banker 1993, Pastor et al. 1995, and Kittelsen 1997). Nevertheless, for simplicity, we focus on Banker's F -test in this paper. However, we stress that the proposed approach can be used in combination with alternative test procedures as well.

One problem in applying the F -test to the above problem is that, as discussed above, the estimators for the outliers are inconsistent. However, the hypothesis can be tested using consistent estimates only. Applying the Banker's F -test to the units in the production set gives the following test statistic, using α for the desired level of significance:

$$(3.7) \quad F_{k,p} = \frac{\sum_{j \in D \setminus Q} (1 - \hat{q}_{REM,k,j})^2}{\sum_{j \in D \setminus Q} (1 - \hat{q}_{REM,p,j})^2} \sim f_{n-q, n-q, 1-\alpha}.$$

Unfortunately, the number and the identity of the outliers are unknown, and hence (3.7) can not be computed directly. However, if the null hypothesis holds, the outliers receive a unity efficiency score, i.e. $\hat{q}_{REM,k,j} = \hat{q}_{REM,p,j} = 1, j \in Q$, and hence do not interfere with the test statistic, i.e.:

$$(3.8) \quad \frac{\sum_{j \in D \setminus Q} (1 - \hat{q}_{REM,k,j})^2}{\sum_{j \in D \setminus Q} (1 - \hat{q}_{REM,p,j})^2} = \frac{\sum_{j \in D} (1 - \hat{q}_{REM,k,j})^2}{\sum_{j \in D} (1 - \hat{q}_{REM,p,j})^2} \quad p \geq k \geq q.$$

In addition, if $k \geq q$, $f_{n-q, n-q, 1-\alpha}$ can be used as an upper bound for $f_{n-k, n-k, 1-\alpha}$. Obviously, this lowers the rejection probability, and hence α is an upper bound for the true level of

Starting from a large enough value $p \geq k$, k can be decreased until the test statistic reaches the critical value corresponding to the desired level of significance (α).

$$k^*(\alpha) = \min_k \left\{ k \mid F_{k-1,p} \geq f_{n-k, n-k, 1-\alpha} \geq F_{k,p} \right\}.$$

Unfortunately, solving (3.2) can be laborious, because it can require the evaluation of a large number of subsets of observations. Fortunately, using duality theory, an alternative formulation can be derived that substantially reduces computational complexity.

Model (2.4) selects the convex combination of observations in the data set that minimizes efficiency. The dual formulation of this problem is well established (e.g. Banker et al. 1984). That formulation selects the linear hyperplane enveloping all observations that maximizes efficiency, i.e.:

$$(3.10) \quad \hat{q}'_{BCC,j} = \max_{u,v,w} \{y_j w - u \mid x_j v = 1; Yw - u - Xv \leq 0; v, w \geq 0\}.$$

Model (3.2) excludes maximally k DMUs from the data set to maximize minimum efficiency. In terms of the dual formulation, that corresponds to maximizing efficiency relative to a linear hyperplane that envelops all but maximally k observations. This maximum can be computed by solving the following Mixed Integer Linear Programming problem:

$$(3.11) \hat{q}'_{REM,j} = \max_{u,v,w,I} \{y_j w - u \mid x_j v = 1; I^T (Yw - u - Xv) \leq 0; I_j = 1; I_{l \neq j} \in \{0,1\}; I^T e = n - k; v, w \geq 0\}.$$

Modern solvers should easily solve this problem even in large-scale problems with numerous input-output dimensions and DMUs.

4. Results from Monte Carlo simulations

In the previous section we discussed some asymptotic properties of REM measures. In order to assess the properties of REM measures in finite samples, we performed a series of Monte-Carlo simulations. In these simulations, two groups of units, say $D \setminus Q = \{1, \dots, 100 - q\}$ and $Q = \{100 - q + 1, \dots, 100\}$, employing different production technologies were included in the same sample, assuming the analyst is not able to identify the two groups *a priori*. The following model formally describes the data generating processes:

$$(4.1) \quad y_j = \begin{cases} |100 + 10x_j - u_j| & j \in D \setminus Q \\ |100 + 20x_j - u_j| & j \in Q \end{cases},$$

$$x_j \sim |N(10, 2)| \quad j \in D,$$

$$u_j \sim |N(0, 10)| \quad j \in D.$$

Inputs follow a normal distribution², and inefficiency terms (u_j) follow a half-normal distribution; i.e. they are the absolute values of random variables with a zero mean normal distribution. Output was computed by first calculating the frontier output, i.e. $(100 + 10x_j)$ for $j \in D \setminus Q$ and $(100 + 20x_j)$ for $j \in Q$, and subsequently subtracting inefficiency.

The DMUs of group Q use a production technology that is superior to that of the DMUs of $D \setminus Q$, and therefore represent outliers for the latter. We considered three different numbers of outliers: $q = 0$, $q = 1$ and $q = 10$.

We compared three models: the standard BCC model (2.4), the REM model (3.2), and the 'ideal' BCC model run after excluding all outliers. For the REM model, we selected the maximum number of exclusions by means of the empirical test procedure discussed in section 3, with $\alpha = 0.05$ and $p = 50$.

In total, 1000 experiments were conducted. Each experiment consisted of generating a set of artificial data from the above data-generating process, and computing efficiency estimates for each data point by using each of the three models. Next, these estimates were used to gauge the sampling distribution properties of the competing estimators. We considered three sample statistics: bias (BIAS), standard deviation (STD), and root mean squared error (RMSE). These statistics were computed as follows:

² Input and output are truncated at zero by using absolute values, so as to prevent negative values.

$$(4.2) \quad BIAS = \frac{1}{n} \sum_{j=1}^n (\hat{q}_j - q_j);$$

$$(4.3) \quad STD = \frac{1}{n} \sum_{j=1}^n (\hat{q}_j - \bar{\hat{q}})^2;$$

$$(4.4) \quad RMSE = \frac{1}{n} \sum_{j=1}^n (\hat{q}_j - q_j)^2$$

Table 4.1 displays the average values (computed over the 1000 replications) for these sample statistics³.

Table 4.1: Simulated sample statistics

		BCC	REM	Ideal
q =0	BIAS	0.016 (0.003)	0.027* (0.006)	0.016 (0.003)
	STD	0.031 (0.016)	0.031 (0.017)	0.031 (0.016)
	RMSE	0.034 (0.016)	0.043* (0.015)	0.034 (0.016)
q =1	BIAS	-0.366 (0.030)	0.012* (0.074)	0.015* (0.006)
	STD	0.063 (0.045)	0.034* (0.047)	0.033* (0.044)
	RMSE	0.397 (0.035)	0.056* (0.078)	0.036* (0.043)
q =10	BIAS	-0.385 (0.019)	-0.110* (0.176)	0.016* (0.005)
	STD	0.079 (0.053)	0.059* (0.042)	0.043* (0.051)
	RMSE	0.429 (0.026)	0.164* (0.167)	0.045* (0.048)

In the first setup including no outliers, the REM procedure had larger bias than the BCC model (the BCC and the 'ideal' model are identical in this case) by wrongly excluding influential

³ We use an asterisk (*) to denote a statistic that differs significantly from the corresponding BCC statistic at a 1 percent significance level, using a paired t-test.

DMUs. However, in the data sets containing outliers, REM offered substantially more accurate estimates relative to the standard model. In fact, the REM estimator comes relatively close to the 'ideal' solution with perfect knowledge of the identity of the outliers. Notice that property (2.5) ($\hat{q}_{BCC,j} \geq q_j$) does not hold for the standard BCC model in the case of outliers, so bias can take also negative values. The negative bias of the REM estimator in the setup with ten outliers was due to the fact that the F-test frequently selected a maximum number of exclusions that was smaller than the number of outliers. Still, REM performs relatively well compared to the standard BCC. This Monte Carlo experiment clearly suggests that the REM procedure is a useful for dealing with outliers in DEA. Nevertheless, more testing is required in different types of setups to study the behavior of the alternative efficiency estimators.

5. Concluding remarks

Excluding outliers from the data set can restore the statistical consistency of DEA estimates. Unfortunately, the identity of the outliers is typically unknown, and there is no unambiguous technique for detecting outliers in DEA. However, simply excluding the most influential observations also gives consistent estimates, provided the number of exclusions exceeds the number of outliers. The smaller the number of exclusions (provided it exceeds the number of outliers), the better finite sample performance. That smallest number can be found by using an empirical test procedure.

Monte-Carlo simulations suggest that this REM approach can substantially improve the estimation relative to the standard BCC model, and in addition that REM can come close to the 'ideal' solution of identifying and excluding all outliers.

Future research could focus on systematically analyzing the performance of the REM model, and comparing it with stochastic DEA models, taking potential specification errors for the latter into account.

In addition, the choice of the specification test may prove critical to the effectiveness of REM. However, relatively little is known about the finite sample performance of the existing tests. Moreover, the existing evidence suggests that these tests are sensitive to finite sample bias and dependence. Therefore, efforts should be directed at gathering systematic evidence on the finite sample performance of existing specification tests, and at designing new ones.

References

- Aigner, D., C.A.K. Lovell, and P. Schmidt (1977), 'Formulation and estimation of stochastic frontier production models', *Journal of Econometrics* 6 (1), 21-37.
- Banker, R. D. (1993): 'Maximum Likelihood, Consistency and Data Envelopment Analysis: A Statistical Foundation', *Management Science* 39 (10), p. 1265 - 1273
- Banker, R.D., A. Charnes, and W. W. Cooper (1984): 'Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis', *Management Science*, Vol. 30, p. 1078 - 1092
- Charnes, A., W. Cooper, and E. Rhodes (1978): 'Measuring the Efficiency of Decision Making Units', *European Journal of Operational Research* 2, p. 429 - 444.

- Charnes, A., S. Haag, P. Jaska, and J. Semple (1992) 'Sensitivity of efficiency classifications in the additive model of Data Envelopment Analysis', *International Journal of Systems Science*, vol. 23, pp. 789-798.
- Cooper, W.W., Z.M. Huang, V. Lelas, S. Xi and O.B. Olesen (1998) 'Chance constrained programming formulations for stochastic characterizations of efficiency and dominance in DEA'. *Journal of Productivity Analysis* 9 (1), pp. 53-79.
- Farrell, M. J. (1957): 'The Measurement of Productive Efficiency', *Journal of Royal Statistical Society, Series A*, p. 253 - 290
- Gong, L., and B. Sun (1995) 'Efficiency measurement of production operations under uncertainty', *International Journal of Production Economics* 39, 55-66.
- Gstach, D. (1998) 'Another Approach to Data Envelopment Analysis in Noisy Environments: DEA+', *Journal of Productivity Analysis* 9 (2), pp. 161-176
- Kittelsen, S.A.C. (1997), Monte Carlo simulations of DEA efficiency measures and hypothesis tests, Paper presented at the 5th European Workshop on Efficiency and Productivity Measurement, Copenhagen, 10-10-1997)
- Korostlev, A., L. Simar, and A.B. Tsybakov (1995a): Efficient Estimation of Monotone Boundaries, *The Annals of Statistics*, 23, p. 476-489
- Korostlev, L. Simar, and Tsybakov (1995b): On Estimation of Monotone and Convex Boundaries, *Pub. Inst. Stat. Univ. Paris*, XXXIX, 1, p. 3-18
- Land, K., C., A. K. Lovell, and S. Thore (1994) 'Chance-Constrained Data Envelopment Analysis'. *Managerial and Decision Economics* 14, 541-554.
- Meeusen, W., and J. van den Broeck (1977) 'Efficiency estimation from Cobb-Douglas production functions with composite error', *International Economic Review* 18, 435-444.
- Olesen, O. B., and N. C. Petersen (1995) 'Chance Constrained Efficiency Evaluation'. *Management Science* 41, 442-457.
- Pastor, J.T., J. L. Ruiz, and I. Sirvent (1995) 'A statistical test for nested radial DEA models'. Working Paper, Departamento de Estadística e Investigación Operativa, Universidad de Alicante, 03071-Alicante, Spain.
- Post, G.T. (1997) 'Performance benchmarking in stochastic environments using Mean-Variance Data Envelopment Analysis'. Rotterdam Institute for Business Economics Studies (RIBES) Report R9701/O, currently reviewed by *Operations Science*.
- Wilson, P. (1995): 'Detecting Influential Observations in Data Envelopment Analysis', *Journal of Productivity Analysis* 6, pp. 27 - 45.
- Zhu, J. (1996) 'Robustness of the efficient DMUs in Data Envelopment Analysis', *European Journal of Operations Research* 90, 451-460.