European Journal of Operational Research 198 (2009) 647–654

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Interfaces with Other Disciplines

Testing procedures for detection of linear dependencies in efficiency models

Antonio Peyrache [a,*], Tim Coelli [b]

[a] Department of Economic Sciences, "La Sapienza" University, Via Avezzano, 1, 00011 Guidonia (RM), Italy
[b] School of Economics, University of Queensland, Brisbane, Australia

ARTICLE INFO

Article history:
Received 8 February 2008
Accepted 26 August 2008
Available online 2 September 2008

Keywords:
Data envelopment analysis
Correlation
Independence
Hypothesis test
Aggregation

ABSTRACT

The validity of many efficiency measurement methods rely upon the assumption that variables such as input quantities and output mixes are independent of (or uncorrelated with) technical efficiency, however few studies have attempted to test these assumptions. In a recent paper, Wilson (2003) investigates a number of independence tests and finds that they have poor size properties and low power in moderate sample sizes. In this study we discuss the implications of these assumptions in three situations: (i) bootstrapping non-parametric efficiency models; (ii) estimating stochastic frontier models and (iii) obtaining aggregate measures of industry efficiency. We propose a semi-parametric Hausmann-type asymptotic test for linear independence (uncorrelation), and use a Monte Carlo experiment to show that it has good size and power properties in finite samples. We also describe how the test can be generalized in order to detect higher order dependencies, such as heteroscedasticity, so that the test can be used to test for (full) independence when the efficiency distribution has a finite number of moments. Finally, an empirical illustration is provided using data on US electric power generation.

1. Introduction

The measurement of technical efficiency has been the subject of many studies since the pioneering work of Farrell (1957). Most of these studies have made the implicit assumption that the degree of technical inefficiency of a firm is independent of the inputs (and output mixes) of the firm.[1] However, there are various reasons why this assumption may be incorrect. For example, Wilson (2003) notes that in some instances big firms may have access to better managers and hence are more likely to perform better. Furthermore, Schmidt and Sickles (1984) argue that if a firm knows its level of technical inefficiency this should affect its input choices, creating a potential dependence between the input vector and the efficiency term.

Wilson (2003) surveys a number of the independence tests that could be used to test the independence hypothesis in the context of efficiency measurement. His motivation essentially relates to the fact that if independence can be assumed, one can implement a much simpler bootstrapping methodology to construct confidence intervals for efficiency estimates derived using data envelopment analysis (DEA). He conducts a Monte Carlo experiment to investigate the small sample properties of four independence testing procedures (two bootstrap-based tests and two rank-based tests) and finds that they all have incorrect size properties and poor power properties when the sample size is not large ($n = 70$) and the degree of correlation (ρ) is moderate ($\rho \leqslant 0.4$), with the rank-based tests not performing as well as the bootstrap tests.

In this study we deviate from the Wilson (2003) study two important ways. First, we discuss two additional situations in which independence information is valuable – namely stochastic frontier models and aggregation of efficiency scores. Secondly, we focus our attention on the hypothesis of uncorrelation (no linear dependence) as opposed to independence. The advantage of testing this weaker condition is that we can produce testing procedures which are easy to implement, and (as we show in our Monte Carlo experiment) have correct size and much stronger power relative to the independence tests. Of course the downside is that the uncorrelation test cannot identify non-linear relationships. However, in the event that the null hypothesis of uncorrelation is rejected, one can also conclude that the null hypothesis of independence is also rejected. Thus providing a valuable pre-test procedure if independence is the hypothesis of interest.

In this study we discuss three important contexts in which these properties play a fundamental role. *First*, in stochastic frontier models (SFM) an uncorrelation assumption is needed for one to conclude that the corrected ordinary least squares (COLS) estimator provides consistent estimates of the slope parameters (Kumbhakar and Knox Lovell, 2000). If correlation between the efficiency term and the regressors arise, we have an endogeneity

* Corresponding author. Tel.: +39 328 96 11863.
E-mail address: apeyrache@msn.com (A. Peyrache).

[1] This statement assumes output oriented technical efficiency measures are being estimated. In the event that one is alternatively estimating input oriented technical efficiency measures, the output levels and the input mixes are the relevant variables. This is explained further in the discussion below.

0377-2217/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2008.08.014

problem. Furthermore, maximum likelihood estimation (MLE) cannot be used when correlation exists because the increased number of parameters in the model gives rise to identification problems. *Second*, in the aggregation of Farrell type efficiency measures (for example, see Fare and Grosskopf, 2004; Fox, 2004) the monotonicity property[2] of the aggregate industry efficiency indexes holds if and only if the uncorrelation assumption is satisfied. The failure of the monotonicity property gives rise to the so called Fox Paradox, where one can find that individual efficiency scores can all increase but the measure of overall industry efficiency decreases. Therefore this paradox can be interpreted as an example of the failure of the uncorrelation assumption. *Third*, the uncorrelation assumption is a necessary condition for independence and this last one is used in non-parametric frameworks to justify the use of univariate kernel methods for the estimation of the efficiency distribution (Wilson, 2003; Daraio and Simar, 2005). If independence fails one has to estimate a multi-dimensional density function, leading to the well known curse of dimensionality problem (Efron and Tibshirani, 1993).

The remainder of this paper is organized into sections. In Section 2 we define the production technology and introduce formal definitions of independence and uncorrelation. Some aggregation issues and the relations between the uncorrelation assumption and the monotonicity property are discussed in Section 3. In Section 4 the impact of the failure of the uncorrelation assumption on stochastic frontier models is explicitly discussed. In Section 5 we introduce some statistical procedures to test for uncorrelation and homoscedasticity. Finally, in Section 6 we conduct a Monte Carlo experiment and provide an empirical illustration of the problems discussed using data on the US electricity power generation industry. Some concluding remarks are then provided in the final section.

2. The technology plus some definitions

2.1. Stochastic representation of technology

Consider the density function $f(\mathbf{x}, \mathbf{y}) \geqslant 0$, where $\mathbf{x} \in R^k$, $\mathbf{y} \in R^m$ are the input and the output vectors and $\int_{R^{k+m}} f(\mathbf{x},\mathbf{y})\,d\mathbf{x}\,d\mathbf{y} = 1$, where $\mathbf{x}$ and $\mathbf{y}$ assume non-negative values. We define the support of the density function as

$$T = \{(\mathbf{x},\mathbf{y}) \in R^{k+m} : \ f(\mathbf{x},\mathbf{y}) > 0\}$$

and its boundary as an intersection between sets

$$\vartheta T = \{[T \cap cl(\overline{T})] \cup [cl(T) \cap \overline{T}]\},$$

where $\overline{T}$ is the compliment of T and $cl(\cdot)$ is the closure operator. In production economics we refer to the first set as the *Production Set* and to the boundary as the *Production Frontier*. The following regularity conditions (Kumbhakar and Knox Lovell, 2000; Fare et al., 1994) are commonly used in production economics: (A1) *no free lunch*: if $f(\mathbf{x}, \mathbf{0}) > 0$ and $f(\mathbf{0}, \mathbf{y}) > 0$ then $\mathbf{y} = \mathbf{0}$; (A2) the *Production Set is Closed*: for a succession of points $(\mathbf{x}_n, \mathbf{y}_n) \to (\mathbf{x}, \mathbf{y})$, if $f(\mathbf{x}_n, \mathbf{y}_n) > 0\ \forall n \in N$ then $f(\mathbf{x}, \mathbf{y}) > 0$; (in essence, this states that the frontier belong to the production set)[3]; (A3) the *Production Set is bounded*: for each $\mathbf{x} \in R^k_+$ exist $\mathbf{y}$: $f(\mathbf{x}, \mathbf{y}) = 0$; (A4) *strong disposability*: if $f(\mathbf{x}_0, \mathbf{y}_0) > 0$ then $f(\mathbf{x}_1, \mathbf{y}_1) > 0$ for each $(-\mathbf{x}_1, \mathbf{y}_1) \leqslant (-\mathbf{x}_0, \mathbf{y}_0)$; (A5) *convexity*: if $f(\mathbf{x}_1, \mathbf{y}_1) > 0$ and $f(\mathbf{x}_2, \mathbf{y}_2) > 0$ then $f[\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2, \alpha\mathbf{y}_1 + (1-\alpha)\mathbf{y}_2] > 0\ \forall 0 \leqslant \alpha \leqslant 1$.

These are pure statistical restrictions on a stochastic data generating process (DGP) represented by the density function $f(\mathbf{x}, \mathbf{y})$. In what follows we assume that assumptions A1–A4 hold. In addition, we assume the following regularity condition on the DGP (Daraio and Simar, 2005):

- *Random Sample*: the sample observations $(\mathbf{x}_i, \mathbf{y}_i)$, $i = 1,\ldots,n$ are realizations of identically and independently distributed random variables $(\mathbf{X}, \mathbf{Y})$ which have probability density function $f(\mathbf{x}, \mathbf{y})$.

2.2. Average technical efficiency, independence, uncorrelation and homoscedasticity

Let us consider the output oriented radial measure of efficiency $\theta = \min\{\theta : f(\mathbf{x}, \frac{\mathbf{y}}{\theta}) > 0\}$. Before we proceed, it is useful to explicitly show that it is possible to calculate the efficiency distribution from the original joint density function $f(\mathbf{x}, \mathbf{y})$. An easy way to calculate the marginal distribution of efficiency is via the method of cylindrical coordinates (Simar and Wilson, 2000). The cylindrical coordinates of a point $(\mathbf{x}, \mathbf{y})$ are $(\tau, \boldsymbol{\eta}, \mathbf{x})$ where $\tau = \sqrt{\mathbf{y}'\mathbf{y}}$ and $\tan\eta_j = \frac{y_j}{y_1}\ \forall j = 1, \ldots, m$. The distance between $(\mathbf{x}, \mathbf{y})$ and its efficient radial projection on the frontier can be stated in cylindrical coordinates as $\theta = \frac{\tau(\mathbf{y})}{\tau(\mathbf{y}/\theta)}$. Since a point $(\mathbf{x}, \mathbf{y})$ is fully represented in cylindrical coordinates $(\tau, \boldsymbol{\eta}, \mathbf{x})$ and we have a biunivocal correspondence between τ and θ, we can write it as $(\theta, \boldsymbol{\eta}, \mathbf{x})$. Then the density function can be written as

$$f(\mathbf{x},\mathbf{y}) = f(\theta,\boldsymbol{\eta},\mathbf{x}) = f(\theta \mid \boldsymbol{\eta},\mathbf{x}) f(\boldsymbol{\eta} \mid \mathbf{x}) f(\mathbf{x}). \tag{1}$$

The marginal efficiency distribution can be calculated by integrating the density function (1) with respect to $\mathbf{x}$ and $\boldsymbol{\eta}$:

$$f_\theta(\theta) = \int\int f(\theta,\boldsymbol{\eta},\mathbf{y})\,d\boldsymbol{\eta}\,d\mathbf{x}. \tag{2}$$

The knowledge of the density function (2) allows one to aggregate efficiency, or in fact to determine all the moments of its distribution. We now provide three useful definitions.

Definition 1 (*Independence*). The efficiency distribution is *fully independent* if and only if $f(\theta|\eta, \mathbf{x}) = f(\theta)$. Efficiency is independent from output composition (or *output composition* independence) if and only if $f(\theta|\eta, \mathbf{x}) = f(\theta|\mathbf{x})$. Furthermore, efficiency is independent from the input set (or *input set* independence) if and only if $f(\theta|\eta, \mathbf{x}) = f(\theta|\eta)$.

Definition 2 (*Uncorrelation or linear independence*). The efficiency distribution is *fully uncorrelated* if and only if $E(\theta\ |\eta, \mathbf{x}) = E(\theta)$. Efficiency is uncorrelated with output composition (or *output compositionu*ncorrelation) if and only if $E(\theta|\eta, \mathbf{x}) = E(\theta|\mathbf{x})$. Furthermore, efficiency is uncorrelated with the input set (or *input set* uncorrelation) if and only if $E(\theta|\eta, \mathbf{x}) = E(\theta\ |\eta)$.

Definition 3 (*Homoscedasticity*). The efficiency distribution is homoscedastic if and only if $\mathrm{Var}(\theta_i) = \mathrm{Var}(\theta)$ or if its variance is constant across observations.

Since $\mathrm{Var}(\theta) = E_2(\theta) + E_1^2(\theta)$ homoscedasticity can be rewritten as[4]

$$E_2(\theta \mid \boldsymbol{\eta},\mathbf{x}) + E_1^2(\theta \mid \boldsymbol{\eta},\mathbf{x}) = E_2(\theta) + E_1^2(\theta) \tag{3}$$

From Eq. (3) it is easy to see that a violation of the uncorrelation assumption implies (excluding some minor cases) a violation of the homoscedasticity assumption.

[2] Given a vector of individual values and an aggregate index based on this individual values, the monotonicity property states that if all the individual values increase also the aggregate index have to increase (Balk, 1995).

[3] The closure of the production set (A2) can be also stated (see Daraio and Simar, 2005; Wilson, 2003) in terms of *Positiveness*: the density function is strictly positive on the boundary and is continuous in any direction toward the interior (i.e., the density function is discontinuous on the boundary).

[4] Note that the expectations notation is such that $E_1(y) = E(y^1)$ and $E_2(y) = E(y^2)$.