A PROBABILISTIC APPROACH FOR ASSESSING THE SIGNIFICANCE OF CONTEXTUAL VARIABLES IN NONPARAMETRIC FRONTIER MODELS: AN APPLICATION FOR BRAZILIAN BANKS

Roberta Blass Staub
Departamento de Estudos e Pesquisas, Banco Central do Brasil
Setor Bancário Sul, quadra 3, bloco B
robertha.blass@bcb.gov.br

Geraldo da Silva e Souza
Universidade de Brasília - Departamento de Estatística, Brasil
geraldo.souza@embrapa.br

ABSTRACT

This article presents an empirical application for the Central Bank of Brazil illustrating the use of a nonparametric frontier model relying on a probabilistic definition of the production frontier. The significance of the variable nonperforming loans in productive efficiency is assessed, for a sample of Brazilian banks, using the concepts of conditional and unconditional efficiency measures, in a context where it is not necessary to impose any particular distribution for the production data. The analysis is robust relative to the assumptions of separability.

KEY WORDS. Stochastic frontier. Contextual variables. FDH efficiency measurements. DEA.

1 The views expressed in this work are those of the authors and do not reflect those of the Banco Central or its members.
1. Introduction

A typical nonparametric (deterministic) efficiency analysis where interest is in the significance assessment of contextual variables is based on the two-stage approach. First an efficiency measurement is obtained (usually by DEA - Data Envelopment Analysis) and, in a second step, the estimated efficiency measurement is regressed on some covariates of interest, assuming an underlying parametric model. Recent examples of this procedure are Souza et al. (2006), Souza and Staub (2007), and Hoff (2006). The validity of the analysis relies on the assumptions of separability (Daraio and Simar (2005)). Basically three types of problems arise in the second stage regression: serial correlation, bias, and correlation between residuals and contextual variables. To overcome these drawbacks related to the two-stage approach, alternative methods of analysis have been developed, among them the probabilistic approach of Daraio and Simar (2005). The idea is to define a new concept of production frontier and efficiency measurement based on probability distributions. We contribute to the literature on this subject using this new methodology in an empirical application for Brazilian banks. Our specific interest is to assess, for a sample of Brazilian banks, the significance of nonperforming loans on the Daraio and Simar stochastic measure of efficiency.

As pointed out by Daraio and Simar (2005), a difference between the two-stage approach and the probabilistic formulation is that the first depends on the separability assumption between the contextual variable vector \( Z \) and the input output pair \((X, Y)\), while the probabilistic approach does not. Separability implies that the production frontier does not change with distinct values of the contextual variables (Daraio and Simar (2005)).

The two-stage approach induced by DEA assumes convexity of the production set. The probabilistic formulation of the frontier uses FDH (Free Disposal Hull) efficiency scores which does not require convexity. Another difference between the approaches is that the probabilistic approach is truly nonparametric while in the two-stage approach one needs to specify a parametric model to be able to regress the estimated efficiency measurements on the contextual variables. Typically, the error term is assumed to follow a truncated normal distribution, although other authors as Banker (1993) and Souza et al. (2006) considered other possibilities, as the use of Tobit like models using the exponential and gamma distributions.

Simar and Wilson (2007) also point out that the inference problems with the two-stage approach may be overcome with the use of the bootstrap. On the other hand, Souza and Staub (2007), extending Banker (1993) results, show that the restrictions pointed by Simar and Wilson are not general. Inference in the two-stage approach can be formally justified, relaxing the assumption of identically distributed inefficiencies, in a deterministic univariate production model, under a set of regularity conditions similar to the separability assumptions.

2. The Probabilistic Frontier

In this section we present a probabilistic interpretation of the Farrell-Debreu (Coelli et al. (2005)) efficiency scores. The formulation we are concerned was proposed by Daraio and Simar (2005) and consider a nonparametric frontier model for which external contextual factors (neither outputs nor inputs) may influence the production process. To properly define this frontier it is necessary to use a probabilistic model to explain the data generating process. In this context, it is possible to define a new concept of conditional efficiency and develop a nonparametric estimator. These ideas were developed initially by Cazals et al. (2002). The authors also proposed the order-m methods due to the sensitivity of DEA and FDH to outliers. An empirical use of these order of ideas can be found in Wheelock and Wilson (2003) in their study of efficiency and technical changes in U.S. commercial banks.

The probabilistic approach formulation is based on the following probability function (Simar and Wilson (2005)):

\[
H_{xy}(y, x) = P(Y \geq y, X \leq x).
\]
The authors provide the following interpretation and property:

1. $H_{XY}(y,x)$ gives the probability that a unit operating at input, output levels $(x,y)$ is dominated, i.e., that another unit produces at least as much output while using no more of any input than the unit operating at $(x,y)$;

2. $H_{XY}(y,x)$ is monotone, nondecreasing in $x$ and monotone nonincreasing in $y$.

The support of $H_{XY}(x,y)$ is the attainable set $\Psi$, i.e.,

$$H_{XY}(y,x) = 0 \quad \forall \quad (x,y) \notin \Psi,$$

where

$$\Psi = \{ (x,y) : x \ can \ produce \ y \}.$$

Applying Bayes’ rule to the probability function $H_{XY}(x,y)$ we get

$$H_{XY}(x,y) = P(X \leq x \mid Y \geq y)P(Y \geq y) = F_{X\mid Y}(x \mid y)S_y(y)$$

and

$$H_{XY}(x,y) = P(Y \geq y \mid X \leq x)P(X \leq x) = S_{Y\mid X}(y \mid x)F_X(x).$$

New concepts of efficiency measures can be defined for the input-oriented and output-oriented cases, assuming $S_y(y) > 0$ and $F_X(x) > 0$. For input orientation

$$\theta(x,y) = \inf \{ \theta \mid F_{X\mid Y}(\theta x \mid y) > 0 \} = \inf \{ \theta \mid H_{XY}(\theta x, y) > 0 \}$$

and for output orientation

$$\lambda(x,y) = \sup \{ \lambda \mid S_{Y\mid X}(\lambda y \mid x) > 0 \} = \sup \{ \lambda \mid H_{XY}(\lambda y, x) > 0 \}.$$

Since the support of the joint distribution of $(X,Y)$ is the attainable set, boundaries of $\Psi$ can be defined in terms of conditional distributions.

Comparing with the DEA (Data Envelopment Analysis) measures, there are differences on the interpretation of the efficiency scores in (6) and (7):

1. Input case: $\theta(x,y)$ is the proportionate reduction of inputs (holding output levels fixed) required for a unit operating at $(x,y)$ to achieve zero probability of being dominated;

2. Output case: $\lambda(x,y)$ is the proportionate increase in outputs required for the unit operating at $(x,y)$ to have zero probability of being dominated, holding input levels fixed.

In our empirical work we choose the output orientation since this orientation provides a comparison with previous studies using DEA (Souza et al. (2006)) and for this reason only this case will be treated in details. The input orientation case is analogous and can be seen in Daraio and Simar (2005). The analysis is based on FDH efficiency measures.

3. Estimating the Output Oriented Efficiency Measure

The Farrell-Debreu stochastic output efficiency measure for given levels of input $(x)$ and output $(y)$ is given on (7) and is nonparametrically estimated by

$$\hat{\lambda}_s(x,y) = \sup \{ \lambda \mid \hat{S}_{Y\mid X}(\lambda y \mid x) > 0 \}$$

where

$$\hat{S}_{Y\mid X}(\lambda y \mid x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(y_i > \lambda x_i).$$
It can be shown that
\[
\hat{\lambda}_n(x, y) = \max_{i: x_i \leq x, y_i \geq y} \left\{ \min_{j=1}^{n} \left( \frac{y_j}{y} \right) \right\}
\]  

(10)

and, as noted by Cazals et al. (2002), coincides with the FDH estimator.

The estimated FDH production set is very sensitive to outliers, and consequently, are the estimated efficiency scores. Daraio and Simar (2005) proposed the concept of the robust order-m efficiency measure to overcome this problem since it considers another definition of the benchmark against which units are compared, with the introduction of a new order-m frontier. Here we avoid the outlier problem trimming the data before the analysis.

4. Assessing the Significance of Contextual Variables

To assess the significance of a continuous contextual variable vector \( Z \) on the output oriented efficiency measurement associated to a one input-one output production process, Cazals et al. (2002) proposed a probabilistic frontier model. Daraio and Simar (2005) extended their approach to the multivariate case. Daraio and Simar (2005) discuss only the input oriented framework which we adapt here for output orientation. Conditioning on \( Z = z \), the efficiency measure is defined by

\[
\hat{\lambda}(x, y | z) = \sup \left\{ \lambda \mid S_y(\lambda y | x, z) > 0 \right\},
\]  

(11)

where

\[
S_y(y | x, z) = \text{Prob}(Y \geq y | X \leq x, Z = z).
\]  

(12)

The nonparametric kernel estimate proposed by Daraio and Simar (2005) is

\[
\hat{S}_{Y \mid x, z}(y | x, z) = \frac{\sum_{i=1}^{n} I(x_i \leq x, y_i \geq y)K\left(\frac{z - y_i}{h_y}\right)}{\sum_{i=1}^{n} I(x_i \leq x)K\left(\frac{z - x_i}{h_x}\right)},
\]  

(13)

where \( K(\cdot) \) is the kernel and \( h_y \) is the bandwidth. The bandwidth selection, suggested by Daraio and Simar (2005), is the \( k \) nearest neighbor method with likelihood cross validation described in Silverman (1986), but other methods can also be used.

The smoothing approach is necessary since \( Z \) is a continuous variable. The basic idea is of selecting a bandwidth \( h \) which optimizes the estimation of the \( Z \) density, in the sense of yielding a density estimate which is close to the true density in terms of the Kullback-Leibler information distance.

Plugging in the estimator \( \hat{S}_{Y \mid x, z}(y | x, z) \) in equation (11), we get the conditional FDH efficiency measure for the output oriented case:

\[
\hat{\lambda}_n(x, y | z) = \sup \left\{ \lambda \mid \hat{S}_y(\lambda y | x, z) > 0 \right\}.
\]  

(14)

Simar and Daraio (2005) alert that the asymptotic properties for this estimator have not yet been derived.
5. Empirical Results

We now consider the problem of assessing the significance of nonperforming loans to some measure of efficiency associated with the production frontier of a sample of 94 Brazilian banks. The analysis is of relevance since a positive association would point to the use of efficiency measurements as an indicator of the bank’s performance. Bank efficiency is one important aspect in global market competition. To find out the characteristics that influence efficiency allows to expand the bank market, resulting in social and economic benefits. Souza et al. (2006) investigated this possibility with DEA measurements. We refine their ideas here to investigate the properties of the FDH frontiers in this context.

The production process we considered is the same as the one defined in Souza et al. (2006) and uses a three dimensional output (securities, loans and demand deposits) and three inputs (labor, stock of physical capital - including the book value of premises, equipments, rented premises and other fixed assets, and loanable funds - including transaction deposits and purchased funds). The year of analysis is 2001 and all production variables are measured as indexes relative to a benchmark. The data base used is COSIF, the plan of accounts comprising balance-sheet and income statement items that all brazilian financial institutions have to report to the Central Bank on a monthly basis (see Souza et al. (2006)). The production data has been screened for the presence of outliers.

As already pointed out, the focus is on analysing the influence of the continuous variable nonperforming loans \((q)\) on the efficient measurements. The main routines to compute the conditional and unconditional probabilistic measures of efficiency where gently provided by Professor Simar and Cinzia Daraio. They were implemented using Matlab.

To achieve our goal we compute the estimates \(\hat{\lambda}_n(x, y)\) and \(\hat{\lambda}_n(x, y | q)\). Significant differences in these estimates means the contextual variable do influence the frontier. The bandwidth selection method (k-nearest neighbor) suggested by Daraio and Simar (2005) and required to obtain the nonparametric estimator of \(S(x | y, q)\) was not adequate in the present instance. The number of observations \(k\) provided by the method was either the full sample size or only one observation. As an alternative, we chose to minimize the mean integrated square error (Silverman (1986)). The value of \(h\) obtained using a quadratic kernel was \(h = 0.5308\).

To evaluate the associatation between the frontier and nonperforming loans, firstly we calculated the Spearman rank correlation coefficient between \(\hat{\lambda}_n(x, y)\) and \(\hat{\lambda}_n(x, y | q)\). The value of this coefficient is \(-0.32\), with a p-value of 0.002, indicating a significant negative association, as expected. The Kolmogorov-Smirnov two-sample test of equality also rejects \(\hat{\lambda}_n(x, y) = \hat{\lambda}_n(x, y | q)\) at the 1% level.

To further investigate the association between the frontier and \(q\), and motivated by the work of Souza and Staub (2007), we fitted a gamma distribution for the error component in a regression where the \(\frac{\hat{\lambda}_n(x, y | q)}{\hat{\lambda}_n(x, y)}\) was the dependent variable. Results are reported in Table 2. The shape parameter \(\rho\) is constant and the scale parameter of the distribution (\(\omega\)) is given by

\[
\omega_i = \exp(-\mu_i)
\]

with

\[
\mu_i = z_0 + \beta_0 + \cdots + z_h \beta_h.
\]

The \(z_j\) are realization of the \(z\) and the \(\beta_j\) are parameters to be estimated.

We also found nonperforming loans significant, and as it increases, the level of efficiency decreases, confirming the negative association with the response variable. The other variables included in the regression were bank classes, more specifically, bank nature \((n)\), bank type \((t)\),
bank size \((s)\), bank control \((c)\) and bank origin \((o)\). These covariates are all categorical. The variable \(n\) assumes one of two values (commercial, multiple), the variable \(t\) assumes one of four values (credit, business, bursary, retail), the variable \(s\) assumes one of four values (large, medium, small, micro), the variable \(c\) assumes one of two values (private, public), and the variable \(o\) assumes one of two values (domestic, foreign). We notice here that Souza et al. (2006) did not find \(q\) to be significant, although they found the categorical variables significant.

We should remark that this result can be due to the use of different response variables. The two-stage approach consist on first calculating a nonparametric DEA efficiency measure and then this efficiency score is regressed on some technical effects. It is based on the separability condition between the input/output space and the contextual variables space, distributional assumptions and linearity.

The probabilistic efficiency measure relies on a new definition of the production process. It is described by the joint probability measure of \((X, Y)\) \((H(x, y))\). The support of the joint distribution is the attainable set, consequently, the production frontier can be obtained in terms of the conditional distribution, in the output case given by \(P(Y \geq y | X \leq x)\). The inclusion of the contextual variable is done by conditioning the joint distribution on \(q\). The separability condition is not assumed and it is not necessary to impose linearity nor any probabilistic distribution.

This new characterization of the frontier, and the efficiency measure, allows for the identification of the influence of nonperforming loans on the efficiency level by analysing the differences between the conditional and unconditional measures. It indicates how important is the choice of how to calculate efficiency and suggest us to explore other measures to find out if significance of other effects are masked.

<table>
<thead>
<tr>
<th>Var</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>(t)</th>
<th>p-value</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-2.545</td>
<td>0.365</td>
<td>-6.96</td>
<td>&lt;.0001</td>
<td>-3.271</td>
<td>-1.820</td>
</tr>
<tr>
<td>(n_1)</td>
<td>0.070</td>
<td>0.319</td>
<td>0.22</td>
<td>0.827</td>
<td>-0.564</td>
<td>0.703</td>
</tr>
<tr>
<td>(n_2)</td>
<td>0.070</td>
<td>0.309</td>
<td>0.23</td>
<td>0.822</td>
<td>-0.544</td>
<td>0.683</td>
</tr>
<tr>
<td>(t_1)</td>
<td>-0.001</td>
<td>0.122</td>
<td>-0.01</td>
<td>0.992</td>
<td>-0.242</td>
<td>0.240</td>
</tr>
<tr>
<td>(t_2)</td>
<td>-0.050</td>
<td>0.098</td>
<td>-0.51</td>
<td>0.611</td>
<td>-0.245</td>
<td>0.145</td>
</tr>
<tr>
<td>(t_3)</td>
<td>-0.179</td>
<td>0.193</td>
<td>-0.93</td>
<td>0.356</td>
<td>-0.562</td>
<td>0.204</td>
</tr>
<tr>
<td>(s_1)</td>
<td>0.026</td>
<td>0.143</td>
<td>0.18</td>
<td>0.854</td>
<td>-0.257</td>
<td>0.310</td>
</tr>
<tr>
<td>(s_2)</td>
<td>-0.027</td>
<td>0.108</td>
<td>-0.25</td>
<td>0.801</td>
<td>-0.242</td>
<td>0.187</td>
</tr>
<tr>
<td>(s_3)</td>
<td>-0.093</td>
<td>0.097</td>
<td>-0.96</td>
<td>0.340</td>
<td>-0.287</td>
<td>0.100</td>
</tr>
<tr>
<td>(c_1)</td>
<td>-0.074</td>
<td>0.105</td>
<td>-0.71</td>
<td>0.481</td>
<td>-0.283</td>
<td>0.134</td>
</tr>
<tr>
<td>(o_1)</td>
<td>0.056</td>
<td>0.072</td>
<td>0.77</td>
<td>0.442</td>
<td>-0.088</td>
<td>0.199</td>
</tr>
<tr>
<td>(q)</td>
<td>-0.064</td>
<td>0.021</td>
<td>-3.01</td>
<td>0.003</td>
<td>-0.105</td>
<td>-0.022</td>
</tr>
<tr>
<td>(P)</td>
<td>12.098</td>
<td>1.741</td>
<td>6.95</td>
<td>&lt;.0001</td>
<td>8.641</td>
<td>15.554</td>
</tr>
</tbody>
</table>

Table 1. Parametric model for the regression of the ratio \(\hat{\lambda}_d(x, y | q) / \hat{\lambda}_d(x, y)\) on \(q\), assuming a gamma distribution with shape parameter \(P\)
6. Summary and Conclusions

We introduced the notion of a probabilistic frontier and the companion definitions of input and output measures of efficiency in a stochastic environment. These ideas are due to Daraio and Simar (2005). The method is based on FDH measures of efficiency and is an alternative less restrictive than the two-stage approach, typically used in DEA models, to assess the significance of contextual variables.

In an application to Brazilian banks we show that the probability notion of efficiency is more informative than the standard DEA concept to detect potential performance failure in the system.

A variety of efficiency models have been suggested in the literature. In the banking context, many papers have focused on cost and profit efficiencies. These efficiency models could also be studied using the Daraio and Simar (2005) methodology, and we would expect that results would change depending on the specific contextual variables that are being employed. Future research could focus on comparing the performance of such models and understanding their advantages/disadvantages and in which context they are useful for regulators and bank risk managers.

Another aspect that can still be explored as an extension of the present work is the use of a similar analysis to a panel data and verify not only the variables that influence the production process, but also if time changes occur as well. Institutions that supervise the banking systems are interested in following bank’s performance through time.
References