Statistical Properties of DEA Estimators in Production Frontiers

Roberta Blass Staub

Brasília, December 2006

Supervisor: Prof. Geraldo da Silva e Souza

Thanks to my family, my supervisor, the Central Bank sponsorship,
friends and to each person that supported me during this period.

Contents

Chapter 1

Introduction

In this work is verified how to assess the significance of factors effects in Data Envelopment Analysis (DEA) measures of efficiency. In the theoretical aspect the thesis contributes to the literature extending Banker's theory ([7]) and in the empirical aspect with an application to Brazilian banks comparing asymptotic, bootstrap and probabilistic approaches.

The theoretical part available focus on how to model DEA inefficiencies as dependent of contextual variables by means of a statistical model similar in appearance to inefficiency component specifications in stochastic frontier models. This is an extension for Banker [7].

In [7], Banker demonstrates that for deterministic univariate production models defined by independent identically distributed inefficiencies, the DEA estimator of a production function maximizes the likelihood of the model if the inefficiency density function is monotone decreasing. Banker also shows that the DEA estimator is weakly consistent and that, in large samples, the distributional assumptions of the true inefficiencies are carried to the estimated inefficiencies.

In this work, Souza and Staub [43] contribute relaxing the assumption of identically distributed inefficiencies from Banker [7] and demonstrating the strong consistency of the DEA production function and showing how one can model inefficiencies in a two-stage approach. Using Monte Carlo simulation, it is found opposite arguments to the critics postulated by Simar and Wilson [42] who assert that estimated DEA efficiencies are correlated and consequently inference in the two-stage approach is invalid. The estimated correlations are inspected in small samples, for a univariate production model assuming original inefficiencies uncorrelated. The observed correlations were negligible and in this case, Banker's results remain valid. The theoretical contributions are available in chapter 3 and forthcoming on the International Transactions of Operations Research (ITOR) journal, Souza and Staub [43].

The main objective of the empirical part is to compute efficiency measures for commercial banks in Brazil and to verify the influence of contextual variables on it. Based

on the recent critics postulated by Simar and Wilson [42] it is worthy to elaborate a comparison of the results of different techniques, since Souza and Staub [43] found that their arguments are not valid in all cases, as demonstrated in the theoretical part of this work. The following techniques are used :

- 1) Maximum likelihood in the context of the truncated normal distribution, the exponential distribution and general Tobit models, as well as nonparametric analysis of covariance (Banker [7] and Souza [11] and [17]) ;
- 2) Randomization process in a parametric analysis of covariance ;
- 3) Simple bootstrap with confidence intervals corrected for the bias (Souza [8]) ;
- 4) Simple and double bootstrap with correlation and bias problems correction (Simar and Wilson [42]) ;
- 5) Probabilistic approach that defines a nonparametric frontier model for the production set ([27]).

The first empirical methodology applied uses output oriented DEA measures of technical efficiency to assess the significance of technical effects for brazilian banks in a two-stage context, with parameters estimated by maximum likelihood. It is based on Banker's [7] and Souza's [11] and [17] results. Inference in the two-stage approach, output simple (combined), is justified by the weak consistency of the production function and that the estimated residuals have approximately, in large samples, the same behavior as the original residuals. Souza [11] extended these conclusions to the heteroscedastic case. Under these assumptions, the estimated dependent variables (residuals and DEA measures) will be independent. Considering multiple output models not necessarily associated to a production model, consistency of the efficiency measure still holds, validating the use of these measures in the two-stage approach, but not the residuals.

The thesis contributes to the literature suggesting a collection of statistical models that could be used in a DEA application using maximum likelihood estimation in the context of the truncated normal distribution, the exponential distribution, and general Tobit models, as well as a nonparametric analysis of covariance. They also improve adequacy checking of the models by using the conditional moment test of specification described in Greene [16]. This work is presented in chapter 4, **Assessing the Significance of Factors Effects in Output Oriented DEA Measures of Efficiency : an Application to Brazilian Banks**, published in Souza *et al.* ([44]).

In chapter 5, **Evaluating the Significance of Factors Effects in Output Oriented DEA Measures of Efficiency by a Randomization Process**, is presented not only the analysis of covariance of the DEA measurement, for a one dimensional and 3-dimensional output vector, but also justified its use by a randomization approach, validating the statistical inference of the model. In this case, properties of the DEA measures relative to the production frontier are not explored, but they are considered as indexes to be adjusted to the covariates.

A bootstrap procedure is described and implemented in chapter 6.1, **Simple Bootstrap Bias Corrected Confidence Intervals for Factors Effects of Brazilian**

Banks DEA Efficiency Measures. To verify the consistency of the results obtained in the best inefficiency model of chapter 4, a bias corrected confidence interval is applied to the brazilian banks data set. The bootstrap also allows us to identify the distributional characteristics of the parameters.

In chapter 6.2, **Estimation and Inference in a Double Bootstrap Applied to the DEA Efficiency Measures of Brazilian Banks**, it is focused on measures of technical efficiency based on Data Envelopment Analysis (DEA) for brazilian banks and related the variation observed to covariates of interest. In the two stage approach for the DEA measures, the work on this thesis innovates using for the brazilian banks data set a double bootstrap, and a DEA measure following a gamma distribution, with combined output, so as to compare the results in chapter 4. The technique is implemented with the aim of correcting for the parameters bias and correlation problems (Simar and Wilson [42]). They justify that these problems invalidate most of the two-stage studies already published.

The previous techniques are based on the separability condition between the input/output space and the contextual variables. It means that the frontier is not influenced by these variables. In the next application, the probabilistic approach, it is assumed this assumption is not valid. In this case, contextual variables affect efficiency if they alter the frontier, when the process is conditioned on them. Also another concept of efficiency measure is used, the probabilistic, where the separability condition is not assumed. The probabilistic nonparametric approach was suggested by Daraio and Simar [27], following the lines in Wheelock and Wilson [45]. The results are shown in chapter 7, **A Probabilistic Approach for Brazilian Banks Contextual Variables in Nonparametric Frontier Models**. This context allows to explore the new concepts of conditional efficiency measure and respective non-parametric estimators.

The inputs used in the analysis are labor, capital and loanable funds. Efficiency measurements are computed for a multiple output (securities, loans, and demand deposits) and for a single (combined) output. The technical effects of interest in the analysis are bank nature (multiple and commercial), bank type (credit, business, bursary and retail), bank size (large, medium, small and micro), bank control (private and public), bank origin (domestic and foreign), and non-performing loans. The latter is a measure of bank risk. The data set is described in chapter 2.

Among the aspects analysed in the DEA application there are : the adequacy of statistical distributions, independent identically distributed inefficiencies assumption, asymptotic results, randomization, parameters bias and separability condition. Also Banker [7] results on the nonparametric estimation of production functions in the context of deterministic models are extended.

Bank efficiency evaluation is closely related with financial stability, a theme of primordial concern for all central banks and financial supervisor institutions all over the world. And it could not be different, since the social and financial consequences of bank crisis can be dramatic. Besides, the increasing number of bankrupted institutions in underdeveloped and developed countries alerts for the strong necessity of avoiding financial

problems.

Central banks have the responsibility of assessing systemic risks and preventing systemic crises. In this context, the degree of efficiency of banks is one of the possible ways to indirectly supervise the quality of the administration of a bank. Additionally, it also provides an extra information that ranks the units being analysed, allowing to compare different performances and to identify the related reasons. It is not only a useful tool for central banks, but also for the own institutions that are interested in being competitive and efficient in financial markets.

Besides the high costs financial instability can cause the society, another remarkable consequence is the loss of confidence in the banking system what can cause subsequent bank ruins and affect the whole economy, since it depends on the health of the financial system. As an example, in daily economy can observed that financial international investments on assets and bonds of a country can be quickly affected by 'bad' news in the financial market due to globalization. Besides, long term investments can be postponed or even canceled, affecting economic growth and social benefits.

The possibility that also other countries suffer the consequences of financial crisis from one country, known as contagion, is another point of interest, since financial crisis can cross frontiers due to the possibility of instantaneous transactions, and also because lots of banks have branches in different countries. It indicates that a conjoint preoccupation of public authorities, motivating them to develop analytical tools for measuring health and performance of financial institutions is of fundamental necessity.

Chapter 2

Specification of Inputs and Outputs

The definition of outputs and inputs in banking is controversial. See Colwell and Davis [26], Berger and Humphrey [22] and Campos [14] for an in depth discussion on the matter. As described in Campos [14] basically two approaches are possible - production and intermediation. The production approach considers banks as producers of deposits and loans using as inputs capital and labor. In such a context typically output is measured by the number of deposit accounts and the number of transactions performed. Under the intermediation approach banks function as financial intermediaries converting and transferring financial assets between surplus units and deficit units. Each output is measured in value not in number of transactions or accounts. There is not a unique recommendation on what should be considered as the proper set of inputs and outputs particularly under the intermediation approach.

The intermediation approach is followed and as output a combination of the components of the vector $y = (v_1, v_2, v_3)$, defined by the variables v_1 = securities, v_2 = loans and v_3 = demand deposits, is taken. This output vector is also combined into a single measure, denoted by y_c, representing the sum of the values of v_i. This approach follows along the lines of Leightner and Lovell [36], Sathie [12] and Campos [14]. Although this definition of output is not always in the banking literature, is the most common, as seen in Campos [14]. Notice for example that the usage of demand deposits in the brazilian banking literature also varies. Nakane [10] studying cost efficiency considers it as a covariate in the cost function although its specification in the translog cost function is similar to an output. Silva and Neto [39], also in the context of cost functions, consider demand deposits only as a factor influencing the technical efficiency component in the model.

All production variables, as shown below, are measured as indices relative to a benchmark and are normalized by a measure of size. This approach has the advantage of making the banks more comparable through the reduction of variability and of the influence of size in the DEA analysis.

It can be emphasized here that DEA is quite sensitive to the dimension and composition of the output vector. Tortosa-Ausina [13] provide examples showing that ordering

in DEA efficiency may change substantially with the dimension of y. A single output is the extreme case. The combined measure has the advantage of avoiding spurious DEA measurements resulting from unique bank specializations. The use of combined output also allows the use of the DEA residuals introduced by Banker [7]. In this sense it leads to more robust and less conservative measures of technical efficiencies. The drawback to its use is that it may show some double counting due to the nature of the output components. But the double counting is also present in the multiple output vector. Nonetheless, most banking studies use a multiple output approach and thus the thesis will follow this literature.

The inputs considered are labor (l), the stock of physical capital (k) which includes the book value of premises, equipments, rented premises and equipment and other fixed assets, and loanable funds (f) which include, transaction deposits, and purchased funds.

Typically the product oriented DEA efficiency analysis variables are specified using input and output measured in physical quantities. This is not strictly necessary and does not prevent its use in the intermediation approach even in a production function context. One may work with indexes or proxies reflecting the intensity of usage of each variable (input or output) in the production process. This is the case with the present application. Total output, loanable funds and capital are values. Also, labor costs is found to be a more reliable measure of the intensity of labor usage than the number of employees which was much variable within the year. In this context, indexes are defined to reflect the behavior of the production variables. These indexes were then further normalized by an index of size defined by the number of employees in the end of the period under analysis.

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. This is the same data base used in all studies on the subject dealing with brazilian banking. See for example Nakane [10] and Campos [14]. The classification of banks was provided by the Supervision Department of the Central Bank of Brazil. They use cluster analysis to group banks according to their characteristics. The total number of banks used in the analysis (sample size) is 94.

As pointed out above output and input variables are treated as indexes relative to a benchmark. In this paper the benchmark for each variable, whether an input, an output or a continuous covariate, was chosen to be the median value of 2001. Banks with a value of zero for one of the inputs or the outputs were eliminated from the analysis. Outputs, inputs, and the continuous covariate were further normalized through the division of their respective indexes by an index of personnel intended to be a size adjusting factor. The construction of this index follows the same method used for the other variables, that is, the index is the ratio of the number of employees in December of 2001 by its median value in the same month.

Even after size adjustment some banks still show values out of range either for inputs or outputs. There are some outliers in the data base. This is a problem for DEA applications which is known to be very much sensitive to outliers. To eliminate non-

conforming output and input vectors, a sort of Mahalanobis distance of common use in regression analysis to identify outlying observations is considered. This amounts to identify as outlying observations for which the ith element of the diagonal of the hat matrix $W(W'W)^{-1}W'$ is at least two times its trace. Here $W = (1, Y)$ or $W = (1, X)$ where 1 is a column of ones and Y and X are the matrices of output products and input usage respectively.

The covariates of interest for our analysis - factors likely to affect inefficiency, are nonperforming loans (q), bank nature (n), bank type (t) , bank size (s), bank control(c) and bank origin (o). Nonperforming loans is a continuous variate and it is also measured as a ratio of indices like an input or output. All other covariates are 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). There is a bank (Caixa Econômica Federal - CEF) in the data base that requires a distinct classification due to its nature - variable n. One more level for this bank is introduced. This amounts to add one more level to the factor bank nature n. Dummy variables were created for each categorical variable. They are denoted $n_1, n_2, n_3, t_1, \cdots, t_4, s_1, \cdots, s_4, c_1, c_2$ and o_1, o_2 respectively.

Chapter 3

Two Stage Inference Using DEA Efficiency Measurements in Univariate Production Models

In the paper **Two Stage Inference Using DEA Efficiency Measurements in Univariate Production Models**, Souza and Staub [43] extends Banker [7] results on the nonparametric estimation of production functions in the context of deterministic models. Relaxing the assumption of iid inefficiencies it is shown the strong consistency of the DEA production function and how one can model effects causing inefficiency, in a manner typically used in stochastic frontier models, using a two stage inference procedure. Asymptotic results are inspected in small samples by means of Monte Carlo simulations. An empirical application illustrates the two stage inference procedure fitting a deterministic production model for the major state company responsible for agricultural research in Brazil. Since the focus of this work is on the empirical results for brazilian commercial banks, this last part will not be reproduced.

The main theoretical results providing justification for these procedures are based on the seminal paper of Banker [7] where it is demonstrated, for deterministic univariate production models defined by iid inefficiencies, that the DEA estimator of a production function maximizes the likelihood of the model if the inefficiency density function is monotone decreasing. It is also shown in Banker's paper that the DEA estimator is weakly consistent and that, in large samples, the distributional assumptions imposed on the true inefficiency variables are carried to the empirical (estimated) inefficiencies. If $g(x)$ is the underlying production function, the deterministic model assumes that actual input-output observations (x_t, y_t) satisfy the statistical model $y_t = g(x_t) - \epsilon_t$ where ϵ_t is the inefficiency random variable.

Recently the inference procedures derived from Banker's article have been put in check by Simar and Wilson [42] and Wilson [15] who argue that the correlation among the DEA efficiency measurements are sizable enough to invalidate the two stage procedure carried out under the assumption of independent errors. In other words, p-values

and t-tests will be wrong. Monte Carlo evidence on the contrary is presented here, at least when the data generating process is defined by a deterministic univariate production model. The correlation observed between estimated inefficiency errors associated to theoretical uncorrelated inefficiencies were negligible in all simulations for all sample sizes considered. Also p-values were not much divergent from what one would expect from the asymptotic theory even for a small sample size.

Relaxing the assumption of identically distributed inefficiency errors it is shown in Souza and Staub [43] that Banker [7] results described above remain valid. Minor modifications are necessary on the original proofs to achieve the extension. The new theoretical framework allows one to model the efficiency measurements in a manner similar to the approach considered in stochastic frontier analysis, were the inefficiency component is assumed to be distributed as a truncated normal or as an exponential random variable with the mean being a monotone function of a linear construct defined by a set of covariates affecting efficiency. See Coelli *et al.* [25] and Kumbhakar and Lovell [35]. These results also allow a better foundation for the models used by Banker and Natarajan [19] to estimate contextual variable effects using DEA under the assumption of stochastic frontier errors with measurement errors bounded above.

3.1 The Statistical Model

Consider the DEA production function defined in section 4.1. Suppose that observations (x_j, y_j) satisfy the statistical model $y_j = g(x_j) - \epsilon_j$, where the technical inefficiencies ϵ_j are nonnegative random variables with probability density functions $f_j(\epsilon)$ monotonically decreasing and concentrated on $\mathcal{R}^+$. The inputs x_j are drawn independently from probability density functions $h_j(x)$ with support set contained in K. Inefficiencies ϵ_j and inputs x_j are also independent.

The likelihood function for the statistical model is given by

$$\mathcal{L}(g) = \prod_{j=1}^{n} f_j(g(x_j) - y_j) h_j(x_j)$$

Theorem 1 *Among all production functions defined in K^*, $g_n^*(x)$ maximizes $\mathcal{L}(g)$. Any other production function $g_o(x)$ such that $g_o(x_j) = g_n^*(x_j)$ also maximizes $\mathcal{L}(g)$.*

Proof For any production function $g(x)$, since $g_n^*(x)$ is of minimum extrapolation, $g(x) \geq g_n^*(x)$ in K_i^*. Then $g_n^*(x_j) - y_j \leq g(x_j) - y_j$. Since $f_j(\epsilon)$ decreases with ϵ the result follows.

□

Theorem 2 *Suppose that the sequence of pairs (x_j, ϵ_j) satisfying the statistical model $y_j = g(x_j) - \epsilon_j$ are drawn independently from the product probability density functions $h_j(x) f_j(\epsilon)$ where*

1. *The sequence of input densities $h_j(x)$ satisfies*

$$0 < l(x) \leq \inf_j h_j(x) \leq \sup_j h_j(x) \leq L(x)$$

for integrable functions $l(x)$ and $L(x)$ and x interior to K.

2. *The inefficiency densities $f_j(\epsilon)$ are such that*

$$F(u) = inf_j F_j(u) > 0, \;\; u > 0$$

where

$$F_j(u) = \int_0^u f_j(\epsilon) d\epsilon.$$

Then if x_0 is a point in K^* interior to K, $g_n^*(x_0)$ converges almost surely to $g(x_0)$.

Proof Let $B(v, \delta)$ denote the open ball with center in v and radius δ. Since $g(x)$ is continuous, given $\Delta > 0$ there exists $\delta_0 > 0$ such that $x \in B(x_0, \delta_0)$ implies $g(x) > g(x_0) - \Delta$. Let

$$A(\delta) = \{(x, \epsilon), x \in B(x_0, \delta) \text{ and } g(x) - \epsilon > g(x_0) - \Delta\}.$$

Consider the event $A_j(\delta) = \{(x_j, \epsilon_j) \in A(\delta)\}$. Since the functions $l(x)$ and $L(x)$ are integrable and $g(x) - g(x_0) + \Delta > 0$ on $B(x_0, \delta_0)$ we may choose $0 < \delta < \delta_0$ such that

$$0 < \int_{B(x_0,\delta)} L(x) dx < 1,$$

and

$$0 < p = \int_{B(x_0,\delta)} l(x) F(g(x) - g(x_0) + \Delta) dx < 1.$$

Now let $p_j = P\{(x_j, \epsilon_j) \in A(\delta)\}$. We have

$$\begin{aligned} 1 > \int_{B(x_0,\delta)} L(x) dx &\geq \int_{B(x_0,\delta)} h_j(x) \left(\int_0^{g(x)-g(x_0)+\Delta)} f_j(\epsilon) d\epsilon \right) dx \\ &\geq \int_{B(x_0,\delta)} l(x) F(g(x) - g(x_0) + \Delta) dx, \end{aligned}$$

and it follows that $0 < p \leq p_j < 1$ for every j. By construction $g_n^*(x) \geq \text{Min}_j y_j$. Thus if $(x_j, \epsilon_j) \in A(\delta)$

$$y_j = g(x_j) - \epsilon_j > g(x_0) - \Delta$$

and $g_n^*(x_0) \geq \text{Min}_j y_j > g(x_0) - \Delta$. Then $g(x_0) - g_n^*(x_0) < \Delta$ and

$$P\{g(x_0) - g_n^*(x_0) \geq \Delta\} \leq P\left\{ \left(\bigcup_{j=1}^n A_j(\delta) \right)^c \right\} = P\left\{ \bigcap_{j=1}^n A_j^c(\delta) \right\} \leq (1-p)^n.$$

Strong consistency then follows from the Borel-Cantelli 0-1 law, since $\sum_{n=1}^{+\infty}(1-p)^n < +\infty$.

□

Assumption 2 of Theorem 2 is satisfied for exponential distributions if the scale parameters are bounded away from zero. It will be true for the general gamma family $\Gamma(r_j, \lambda_j) = \lambda^{r_j} x^{r_j - 1} \exp\{-\lambda_j x\} / \Gamma(r_j)$ if the parameters λ_j and r_j are restricted to closed intervals $[a, b]$ with $0 < a < b$. It will be true for the family of half-normal distributions $N^+(0, \sigma_j^2)$ if the sequence σ_j^{-1} is bounded away from zero. It will also hold for positive truncations of the $N(\mu_j, \sigma_j^2)$ if the parameters μ_j and σ_j^2 satisfy $\sigma_j/\mu_j \in [-D; D]$ for some $D > 0$.

Theorem 3 *Suppose that Assumptions 1 and 2 of Theorem 2 are satisfied and that x_j is interior to K for every j. Let M be a subset of the DMUs included in the sample that generates the n production observations. The asymptotic joint distribution of the technical inefficiencies $\epsilon_{nj}^* = g_n^*(x_j) - y_j,\ j \in M$, coincides with the product distribution of the $\epsilon_j,\ j \in M$.*

Proof The following proof mimics Banker [7]. Since $g(x) \geq g_n^*(x)$ we have

$$\epsilon_j = g(x_j) - y_j \geq g_n^*(x) - y_j = \epsilon_{nj}^*.$$

Let E_j be constants and define $A_m = \bigcap_{j \in M} \{\epsilon_j \leq E_j + 1/m\}$. The sequence A_m decreases to $\bigcap_{j \in M} \{\epsilon_j \leq E_j\}$. On the other hand, for every m,

$$\bigcap_{j \in M} \left\{\epsilon_{nj}^* \leq E_j\right\} = \left[\left(\bigcap_{j \in M} \left\{\epsilon_{nj}^* \leq E_j\right\}\right) \bigcap A_m\right] \bigcup \left[\left(\bigcap_{j \in M} \left\{\epsilon_j^* \leq E_j\right\}\right) \bigcap A_m^c\right].$$

Then

$$P\left(\bigcap_{j \in M} \left\{\epsilon_{nj}^* \leq E_j\right\}\right) \leq P\left(A_m\right) + P\left(\left(\bigcap_{j \in M} \left\{\epsilon_{nj}^* \leq E_j\right\}\right) \bigcap A_m^c\right)$$

and therefore

$$P\left(\bigcap_{j \in M} \left\{\epsilon_{nj}^* \leq E_j\right\}\right) \leq P\left(A_m\right) + \sum_{j \in M} P\left(\left\{\epsilon_j - \epsilon_{nj}^* > \frac{1}{m}\right\}\right).$$

Since

$$\epsilon_j - \epsilon_{nj}^* = \epsilon_j - y_j + y_j - \epsilon_{nj}^* = g(x_j) - g_n^*(x_j),$$

let $n, m \to \infty$ to obtain by Theorem 3.2

$$\text{limsup}_{n \to \infty} P\left(\left\{\epsilon_{nj}^* \leq E_j, \forall j \in M\right\}\right) \leq P\left(\left\{\epsilon_j \leq E_j, \forall j \in M\right\}\right).$$

Also,

$$\bigcap_{j \in M} \left\{\epsilon_j \leq E_j\right\} \subseteq \bigcap_{j \in M} \left\{\epsilon_{nj}^* \leq E_j\right\},$$

and hence

$$\text{liminf}_{n\to\infty} P\left(\left\{\epsilon^*_{nj} \le E_j, \forall j \in M\right\}\right) \ge P\left(\{\epsilon_j \le E_j, \forall j \in M\}\right).$$

□

3.2 Statistical Inference

Theorem 3 is basic for statistical inference in the context of the deterministic production model. The following proposition shows how to construct confidence intervals for the production values $g(x_i)$. Joint confidence intervals may be obtained using Bonferroni's method.

Proposition 1 *Under the assumptions of Theorem 2 let $\hat{q}_i$ be such that $P\{\epsilon^*_{ni} \le \hat{q}_i\} = 1-\alpha$. The interval $[g^*_n(x_i), g^*_n(x_i) + \hat{q}_i]$ has asymptotically level $1-\alpha$ for $g(x_i)$.*

Proof Since $g^*_n(x_i) \le g(x_i)$ it follows that

$$g^*_n(x_i) \le g(x_i) - \epsilon_i + \epsilon_i = y_i + \epsilon_i \le g^*_n(x_i) + \epsilon_i.$$

Therefore $0 \le g(x_i) - g^*_n(x_i) \le \epsilon_i$. Let q_i be the quantile of ϵ_i of order $1-\alpha$. Since $\epsilon_i \le q_i$ implies $0 \le g(x_i) - g^*_n(x_i) \le q_i$ it follows that $[g^*_n(x_i), g^*_n(x_i) + q_i]$ has level $1-\alpha$. Since for large n $\epsilon^*_{ni} \sim \epsilon_i$ by Theorem 3.3, the result follows. □

The next two propositions assume iid inefficiencies when the common inefficiency distribution is either exponential or half-normal. These results are due to Banker [7] and here they are refined to include a measure of goodness of fit.

Proposition 2 *Under the assumptions of Theorem 2 suppose that the ϵ_i are iid with a common exponential density $f(\epsilon) = \lambda \exp\{-\lambda\epsilon\}$, $\lambda, \epsilon > 0$. Let M be any subset of DMUs with m elements. Then*

1. *The quantity $2\lambda \sum_{i\in M} \epsilon^*_{ni}$ has, approximately, a chi-square distribution with $2m$ degrees of freedom.*
2. *If M is the complete set of DMUs then*

$$\frac{2\sum_{i=1}^n \epsilon^*_{ni}}{s},$$

where s is the sample standard error of the estimated residuals ϵ^*_{ni}, has, approximately, a chi-square distribution with $2n$ degrees of freedom.

Proof Since the true inefficiencies ϵ_i are iid exponential with parameter λ then $2\lambda \sum_{i\in M} \epsilon_i$ is chi-square with $2m$ degrees of freedom. If M coincides with the sample the distribution will be chi-square with $2n$ degrees of freedom.

Let $F_n(u)$ be the distribution function of the chi-square distribution with $2n$ degrees of freedom. Given $u, v > 0$, since the chi-square densities are uniformly bounded, there exists a constant C such that

$$\mid F_n(u) - F_n(v) \leq C \mid u - v \mid .$$

Let $\hat{F}_n(u)$ be the distribution function of $2\sum_{i=1}^{n} \epsilon_i/s$. Since $\hat{F}_n(u) = F_n(\lambda s u)$ it follows

$$\mid \hat{F}_n(u) - F_n(u) \leq C \mid \lambda s - 1 \mid u.$$

Statements 1 and 2 are then true for the inefficiencies ϵ_i since s is strongly consistent for λ^{-1}. By Theorem 3.3 they will also hold, approximately, for the ϵ^*_{ni}. □

Proposition 3 *Under the assumptions of Theorem 2 suppose that the ϵ_i are iid with a common half-normal density $f(\epsilon) = \left(2/\sqrt{2\pi}\sigma\right) \exp\{-\epsilon^2/2\sigma^2\}$, $\sigma > 0$, $\epsilon > 0$. Let M be any subset of DMUs with m elements. Then*

1. *The quantity $\sum_{i\in M}(\epsilon^*_{ni})^2/\sigma^2$ has, approximately, a chi-square distribution with m degrees of freedom.*
2. *If M is the complete set of DMUs then the quantities*

$$S_1 = \left(\frac{2}{\pi}\right) \frac{\sum_{i=1}^{n} \left(\epsilon^*_{ni}\right)^2}{\left(\bar{\epsilon}^*_n\right)^2}$$

and

$$S_2 = \left(1 - \frac{2}{\pi}\right) \frac{\sum_{i=1}^{n} \left(\epsilon^*_{ni}\right)^2}{s^2}$$

where $\bar{\epsilon}^*_n$ and s^2 are the sample mean and the sample variance of the ϵ^*_{ni}, respectively, have, approximately, a chi-square distribution with n degrees of freedom.

Proof Under the assumptions $\sum_{i\in M} \epsilon_i^2/\sigma^2$ is chi-square with m degrees of freedom. If M coincides with the sample then the distribution is chi-square with n degrees of freedom. Since the mean of the half-normal distribution is $\sigma\sqrt{2/\pi}$, the variance is $(1 - 2/\pi)\sigma^2$, and the chi-square densities are uniformly bounded, Results 1 and 2 are then true for the inefficiencies ϵ_i. By Theorem 3.3 they will also hold for the ϵ^*_{ni}. □

The second statement appearing in Propositions 2 and 3, respectively, are essentially goodness of fit measures and serve the purpose to test if the inefficiencies are iid with the common distribution specified (exponential or half-normal). An alternative test of this hypothesis, with a nonparametric flavor, can be carried out when the underlying hypothesized distribution is assumed to be exponential. This is the Lilliefors test (Conover, 1998) which is a Kolmogorov-Smirnov type statistic. A similar result is not known

to the author for the half-normal distribution. QQ-plots, however, can always be used to inspect departures from both parametric specifications.

The first statement appearing in Propositions 2 and 3, respectively, are used by Banker [7] to assess the difference in efficiencies between two groups M_1 and M_2 of decision making units with m_1 and m_2 elements respectively. If the groups do not differ the ratios $\sum_{i\in M_1}\epsilon^*_{ni}/\sum_{i\in M_2}\epsilon^*_{ni}$ and $\sum_{i\in M_1}(\epsilon^*_{ni})^2/\sum_{i\in M_2}(\epsilon^*_{ni})^2$ will follow the F-distribution with $(2m_1, 2m_2)$ and (m_1, m_2) degrees of freedom, respectively, depending on the assumption imposed on the inefficiency distribution, namely exponential or half-normal. A similar test may be employed to assess the scale of operation in $g(x)$. See Banker and Natarajan (2004).

It should be pointed out that Theorems 2 and 3 allow more flexible parametric specifications for the inefficiencies than those suggested by Propositions 2 and 3. Suppose that $z_0, \ldots, z_l$ are variables we believe to matter in explaining inefficiencies. Following the Coelli, Battese, and Rao (1998) approach to stochastic frontier analysis it can be postulated that

$$\epsilon_i = z_{i0}\delta_0 + \ldots + z_{il}\delta_l + w_i$$

where the δ_j are parameters to be estimated, the z_{ij} are realizations of the z_j and w_i is the truncation of the normal $N(0, \sigma^2)$ at $-\mu_i$. These assumptions are consistent with non-negative truncations of the $N(\mu_i, \sigma^2)$ with $\mu_i = z_{i0}\delta_1 + \ldots + z_{il}\delta_l$. This model may be fitted by maximum likelihood with the ϵ^*_{ni} replacing the ϵ_i. One notices that the mean of the positive truncation of the $N(\mu_i, \sigma^2)$, $\mu_i + \sigma\lambda_i$ and the variance, $\sigma^2\left[1 - \lambda_i(\mu_i/\sigma + \lambda_i)\right]$ where $\lambda_i = \frac{\phi(\mu_i/\sigma)}{\Phi(\mu_i/\sigma)}$, $\phi(.)$ and $\Phi(.)$ being the density and the distribution function of the standard normal respectively, are both monotonic functions of μ_i. The formulation also allows heteroscedasticity. Group comparisons in the context studied in Propositions 2 and 3 can be performed in this more general setting taking some of the z_j to be appropriate dummy variables. Any number of groups is allowed. The same ideas may be applied to the exponential family of densities $\lambda_j \exp\{-\lambda_j t\}$ imposing $\lambda_j = \exp\{-\mu_j\}$. These two families, i.e, the exponential and the truncated normal, as in the stochastic frontier analysis, seem to provide enough flexibility in applications.

The gamma distribution may not be fit by maximum likelihood directly since, typically, some DEA residuals will be zero. This contingency may be resolved adding to ϵ^*_{ni} a positive random variable converging in probability to zero, or using a truncated model at, for example 1/n. These procedures will not destroy the approximation given by Theorem 3.3. In this context one may also fit the gamma density $\lambda_j^p t^{p-1}\exp\{-\lambda_j t\}/\Gamma(p)$ imposing $\lambda_j = \exp\{-\mu_j\}$.

3.3 Monte Carlo Simulations

The objective in this section is to show the Monte Carlo simulation used to illustrate and verify the asymptotic results described in Propositions 2 and 3 and based on Theorems 4 and 5.

To perform the Monte Carlo simulation, consider the Cobb-Douglas production function $y = 100x_1^{0.3}x_2^{0.7}$ where inputs (x_1, x_2) are generated independently from the uniform distributions. Two distributions, the exponential and the half normal, are used to model the inefficiencies. For each of these two distributions it is considered two means (300 and 600) and three sample sizes : $n = 30$, 90 and 150. The simulation process mimics the assumptions set forth in Banker [7].

Two subgroups of $n/2$ DMUs are compared for each sample size n by means of F tests. The process is repeated 1500 times.

The simulation process is defined as follows :

1. Repeat steps [a]-[d] to obtain 1500 samples of n DMUs for which the inefficiency distribution has mean μ.

   [a] Generate the inputs x_{1i} and x_{2i} independently from a uniform distribution in $(47, 57)$ and $(67, 77)$;

   [b] Compute the true output using the Cobb-Douglas production function $g(x_{1i}, x_{2i}) = 100x_{1i}^{0.3}x_{2i}^{0.7}$;

   [c] Generate the technical inefficiencies ϵ_i for the half-normal or the exponential distribution with mean μ. [1]

   [d] Compute the actual output values $y_i = g(x_i) - \epsilon_i$;
2. For each of the 1500 samples of size n compute the DEA technical inefficiencies ϵ_{ni}^* defined in Theorem 3 based on (y_i, x_{1i}, x_{2i}) for $i = 1, \cdots, n$.
3. Given one of the 1500 samples of size n divide it into two subsamples with $m = n/2$ elements each. Compute the appropriate F-statistics for the exponential and the half normal assumptions.
4. The F-statistics should follow, approximately, the $F_{(n,n)}$ distribution when the inefficiencies are exponential and the $F_{(n/2,n/2)}$ distribution when the inefficiencies are half normal.

The evidence from Tables A.15 and A.16, based on the F distribution, is that the empirical quantiles are converging to the theoretical quantiles as expected, considering both distributions. Even for $n = 30$ the theoretical approximations are acceptable. Results seem to be robust relative to the number of DMUs considered in each group and the means of the underlying distributions.

In regard to correlations involving the DEA residuals, no significant values were observed. They seem to mimic the order of magnitude of the correlations generated by the simulated inefficiencies regardless of the distribution generating the data.

[1] To generate a random variate from the exponential distribution with mean 600, generate a random variate from the density $exp\{-x\}$, $x > 0$ and multiply this number by 600. To generate a random variate from the half normal distribution with mean 600, generate a random number w from the uniform distribution and compute $\sqrt{\pi/2} \times 600 \times \Phi^{-1}(\frac{1+w}{2})$ where $\Phi(x)$ is the distribution function of the standard normal.

Chapter 4

Assessing the Significance of Factors Effects in Output Oriented DEA Measures of Efficiency : an Application to Brazilian Banks

The main objective of this paper is to compute measures of technical efficiency based on Data Envelopment Analysis (DEA) for the brazilian banks and to relate the variation observed in these measurements to covariates of interest. This association is investigated in the context of several alternative models fit to DEA measurements of efficiency and DEA residuals. The DEA residuals are derived from a single output oriented DEA measure. They were introduced as a formal tool of analysis in DEA by Banker [7].

Output is measured both as a 3-dimensional vector formed by the variables investment securities, total loans and demand deposits and as a combined index of these variables. The three input sources are labor, capital and loanable funds. The causal factors considered here as affecting efficiency measurements an DEA residuals are bank nature, bank type, bank size, bank control, bank origin and risky loans (nonperforming loans).

The statistical methods used explore Banker [7] and Souza [11] and [17] results.

Several bank studies, among them Eisenbeis *et al.* [29], Sathye [12], Campos [14] and Tortosa-Ausina [13] have considered the use of DEA to measure the relative efficiency of a bank. Typically a DEA context is defined, such as a revenue or cost optimization, input or output orientation, under constant or variable returns to scale, and subsequently analyzed. If additionally an empirical investigation on the association between technical effects and DEA measures is demanded, as in Eisenbeis *et al.* [29], regression is the basic technique used in the analysis. The models suggested in the literature go from the standard analysis of covariance models as suggested in Coelli *et al.* [25], to the Tobit model as in McCarthy and Yaisawarng [37].

Our contribution to this literature is twofold. Firstly we open the possibility of combining output in banking studies which makes the Banker [7] kind of approach viable in a context inherited from a production model. Relative to such models it is possible, besides the assessment of significance of factor effects, to attach measures of error to DEA efficiency measurements. Secondly, even if a deterministic univariate production model is not justifiable one could still make use of a general class of censored models to fit the DEA measurements, whether they are computed in the form of residuals from a production model or simply as a measure of efficiency. In this context, the models we use are similar in appearance to those used in the analysis of a stochastic frontier in a DEA analysis. This is achieved generalizing the Tobit. The distributions other than the normal considered in these extensions are the gamma and the truncated normal. This order of ideas appears in Souza [17] and generalizes Banker and Natarajan [19].

4.1 Data Envelopment Analysis (DEA)

Consider a production process with n production units (banks). Each unit uses variable quantities of p inputs to produce varying quantities of different outputs y. Denote by $Y = (y_1, \cdots, y_n)$ the $s \times n$ production matrix of the n banks and by $X = (x_1, \cdots, x_n)$ the $p \times n$ input matrix. Notice that the element $y_r \geq 0$ is the $s \times 1$ output vector of bank r and x_r is the $p \times 1$ vector of inputs used by bank r to produce y_r (the condition $l \geq 0$ means that at least one component of l is strictly positive). The matrices $Y = (y_{ij})$ and $X = (x_{ij})$ must satisfy : $\sum_i p_{ij} > 0$ and $\sum_j p_{ij} > 0$ where p is x or y.

In our application $p = 3$ and $s = 1$ or $s = 3$ and it will be required $x_r, y_r > 0$ (which means that all components of the input and output vectors are strictly positive).

Definition 1 : The measure of technical efficiency of production of bank o under the assumption of variable returns to scale and output orientation is given by the solution of the linear programming problem $Max_{\phi,\lambda}\phi$ subject to the restrictions :

1. $\lambda = (\lambda_1, \cdots, \lambda_n) \geq 0$ and $\sum_i^n \lambda_i = 1$;
2. $Y\lambda \geq \phi y_o$;
3. $X\lambda \leq x_o$.

In the next part we consider statistical models adequate to the analysis of the optimum values ϕ_0^* of **Definition 1** when covariates are thought to affect them. These models can be viewed as extensions of the univariate case, i.e, when $s = 1$. In this instance it is possible to model the input-output data observations as a production model for which the DEA measurements under certain conditions behave as nonparametric maximum likelihood estimators. These results were originally presented in Banker [7] and are extended in Souza [11].

Suppose that $s = 1$ and that the production pairs (x_i, y_i), $i = 1, \cdots, n$ for the n banks in the sample satisfy the deterministic statistical model

$$y_i = g(x) - \epsilon_i \tag{4.1}$$

where $g(x)$ is an unknown continuous production function, defined on a compact and convex set K. We assume $g(x)$ to be monotonic and concave. The function $g(x)$ also satisfy $g(x_i) \geq y_i$ for all i. The quantities ϵ_i are inefficiencies which are independently distributed nonnegative random variables. The input variables x_i are drawn independently of the ϵ_i.

One can use the observations (x_i, y_i) and Data Envelopment Analysis to estimate $g(x)$ only in the set

$$K^* = \left\{ x \in K; x \geq \sum_{i=1}^{n} \lambda_i x_i, \lambda_i \geq 0, \sum_{i=1}^{n} \lambda_i = 1 \right\}. \tag{4.2}$$

For $x \in K^*$ the DEA production function is defined by

$$g_n^*(x) = sup \left\{ \sum_{i=1}^{n} \lambda_i y_i; \sum_{i=1}^{n} \lambda_i x_i \leq x \right\} \tag{4.3}$$

where the *sup* is restricted to nonnegative vectors λ satisfying $\sum_{i=1}^{n} \lambda_i = 1$.

For each bank o, $g_n(x_0) = \phi_0^* y_0$. This function is a production function on K^*, in other words, is monotonic, concave, $g_n(x_i) \geq y_i$, and satisfies the property of minimum extrapolation, that is, for any other production function $g_u(x)$, $x \in K$, $g_u(x) \geq g_n^*(x)$, $x \in K^*$.

4.2 Statistical Models Adequate to Study Product Oriented DEA Inefficiencies

We begin our discussion here assuming $s = 1$. It is shown in Banker [7] that $g_n(x)$ is weakly consistent for $g(x)$ and that the estimated residuals :

$$\epsilon_i^* = (1 - \phi_i^*) y_i \tag{4.4}$$

have approximately, in large samples, the same behavior as the ϵ_i. Souza [11] shows that the same results hold under conditions that do not rule out heteroscedasticity. These results validate the use of the DEA residuals or inefficiencies, or even the DEA measurements themselves, as dependent variables in regression problems since under the assumptions of the deterministic model they will be independent.

Banker [7] discusses two distributions for the ϵ_i (assumed to be iid random variables) consistent with the asymptotic results cited above : the exponential and the half normal.

Souza [11] extends the discussion to the exponential and truncated normal relaxing the iid assumption. These more general models allow the use of typically stochastic frontier methods in the DEA analysis.

One may argue that the use of distributions like the exponential or the truncated normal are not totally adequate since in any particular application of DEA some residual observations will be exactly zero. This leads one naturally to the consideration of censored models to describe the stochastic behavior of the DEA residuals.

Let $z_0, \cdots, z_b$ be variables (covariates) we believe to affect inefficiency. Based on Souza [11] results the following two statistical models can be used to fit the inefficiencies ϵ_i^* under the assumptions of the deterministic model.

Firstly one may postulate the exponential density

$$\lambda_i exp(-\lambda_i \epsilon)$$

where $\lambda_i = exp(-\mu_i)$ with

$$\mu_i = z_{0i}\beta_0 + \cdots + z_{bi}\beta_b. \tag{4.5}$$

The z_{ji} are realizations of the z_j and the β_j are parameters to be estimated.

Secondly one may consider the model $\epsilon_i^* = \mu_i + w_i$ where w_i is the truncation at $-\mu_i$ of the normal $N(0, \sigma^2)$. This model is inherited from the analysis of stochastic frontiers of Coelli *et al.* [25] and is equivalent to truncations at zero of the normals $N(\mu_i, \sigma^2)$.

For the exponential distribution the mean of the *ith* inefficiency error is $exp(\mu_i)$ and the variance $exp(2\mu_i)$. For the truncated normal the mean is

$$\mu_i + \sigma\xi_i \tag{4.6}$$

and the variance

$$v_i = \sigma^2 \left(1 - \xi_i \left(\frac{\mu_i}{\sigma} + \xi_i\right)\right) \tag{4.7}$$

where

$$\xi_i = \frac{\phi(\mu_i/\sigma)}{\Phi(\mu_i/\sigma)}$$

$\phi(.)$ and $\Phi(.)$ being the density function and the distribution function of the standard normal, respectively.

In both models the mean and the variance are monotonic functions of μ_i and thus both specifications allow monotonic heteroscedasticity.

A censored model discussed in Souza [17] that could also be used impose the assumption that the ϵ_i^* satisfies the statistical model

$$\epsilon_i^* = \begin{cases} w_i, & if \;\; w_i > 0 \\ 0, & if \;\; w_i \leq 0, \end{cases}$$

where $w_i = \mu_i + u_i$ the u_i being iid normal errors with mean zero and variance σ^2. This is the Tobit model of McCarthy and Yaisawarng [37]. An extension allowing heteroscedasticity can be introduced assuming that the variance σ^2 is dependent on i and on some set of observables l_i, in other words, $\sigma_i^2 = \exp\{(1, l_i')\zeta\}$, where the parameter vector ζ is unknown. In our application this dependency will be on bank size.

The Tobit model is adequate when it is possible for the dependent variable to assume values beyond the truncation point, zero in the present case. McCarthy and Yaisawarng [37] argue that this is the case in the DEA analysis. Their wording on this matter is as follows. It is likely that some hypothetical banks might perform better than the best banks in the sample. If these unobservable banks could be compared with a reference frontier constructed from the observable banks, they would show efficiency scores less than unity (over efficiency). This would lead to a potential non positive residual.

Clearly the Tobit could also be defined for the efficiency measurements ϕ_i^* in which case the truncation point would be one. We would have

$$\phi_i^* = \begin{cases} w_i, & if \;\; w_i > 1 \\ 0, & if \;\; w_i \leq 1. \end{cases}$$

Maybe a more reasonable assumption in the context of the Tobit model is to allow only for positive over efficiencies. In this case the distributions that readily come to mind to postulate for w_i are the truncation at zero of the normal $N(\mu_i, \sigma^2)$ and the gamma with shape parameter constant P and scale λ_i.

The standard technique to analyze all these models is maximum likelihood. The likelihood functions to be maximized with respect to the unknown parameters are defined as follows.

For the exponential distribution is

$$L(\delta) = \prod_{i=1}^{n} \lambda_i \exp\{-\lambda_i \epsilon_i^*\}.$$

For the truncated normal is

$$L(\delta, \sigma) = \prod_{i=1}^{n} \frac{\phi\left(\frac{\epsilon_i^* - \mu_i}{\sigma}\right)}{\sigma \Phi\left(\frac{\mu_i}{\sigma}\right)}$$

where $\phi(.)$ is the density of the standard normal and $\Phi(.)$ its distribution function.

For the heteroscedastic Tobit model with censoring point at $a = 0$ or $a = 1$ is

$$L(\delta, \zeta) = \prod_{i:y_i^*=a} \Phi\left(\frac{a-\mu_i}{\sigma_i}\right) \prod_{i:y_i^*>a} \frac{1}{\sigma}\phi\left(\frac{y_i^*-\mu_i}{\sigma_i}\right)$$

where $y_i^* = \epsilon_i^*$ or $y_i^* = \phi_i^*$.

For the Tobit with censoring defined by a truncated normal is

$$L(\delta, \sigma) = \prod_{i:y_i^*=a} \frac{\Phi\left(\frac{a-\mu_i}{\sigma}\right) - \Phi\left(\frac{-\mu_i}{\sigma}\right)}{\Phi\left(\frac{\mu_i}{\sigma}\right)} \prod_{i:y_i^*>a} \frac{1}{\sigma}\frac{\phi\left(\frac{y_i^*-\mu_i}{\sigma}\right)}{\Phi\left(\frac{\mu_i}{\sigma}\right)}.$$

For the Tobit with censoring defined by a gamma distribution, let $\Gamma(.)$ denote the gamma function and let $G_p(.)$ denote the distribution function of the gamma distribution with shape parameter P and unit scale. The likelihood is

$$L(\delta, p) = \prod_{i:\phi_i^*=1} G_p(\lambda_i) \prod_{i:\phi_i^*>1} \frac{\lambda_i^p(\phi_i^*)^{p-1}\exp\{-\lambda_i\phi_i^*\}}{\Gamma(p)}.$$

Some of the results in Banker [7] and Souza [11] can be extended to multiple output models not necessarily associated to a production model. Consistency of the ϕ_i^* is one of them. See Kneip *et al.* [32] and Banker and Natarajan [18] and [19]. This suggests that with the exception of the censoring at zero case and the models for DEA residuals all approaches are viable for multiple outputs since in large samples the DEA measurements ϕ_i^* will behave as in random sampling.

Another class of models that can be used in any instance is defined by the class of analysis of covariance models as suggested in Coelli *et al.* [25]. Here we apply a nonparametric version of the analysis of covariance taking as responses the rankings r_i of the observations on the variables under investigation (Conover [9]). In other words we also use the model

$$r_i = z_{0i}\delta_0 + \cdots + z_{bi}\delta_b + u_i \tag{4.8}$$

where the u_i are independent $N(0, \sigma^2)$ non observable errors. This model shows approximate nonparametric properties.

4.3 Data Analysis

We begin the discussion in this section with Tables A.1 to A.4 which show basic statistics for DEA measures. DEA analysis was carried out using the software Deap 2.1

(Coelli *et al.* [25]). Entries in Tables A.1 to A.3 relate to the behavior and the association of the DEA measures of efficiency considered here and Table A.4 is a runs test of randomization (Wonnacott and Wonnacott [46]). We do not see evidence from this table against the assumption of independent observations. Table A.1 refers to $(\phi_i^*)^{-1}$ when the output is y_c, i.e, combined. Table A.2 refers to the same variable when the output is trivariate. Table A.3 presents a matrix of Spearman rank correlations between the three responses of interest - DEA residuals and DEA measurements computed assuming combined and multiple output. The rank correlations seem to point to differences in the analysis with each variable. Although efficiency measurements computed considering the multiple output are much larger than the corresponding measurements for the combined output the ordering induced by the two measures show a reasonable agreement. For bank size and bank nature the averages of both measurements point to the same direction. Commercial banks dominate multiple banks and small and micro banks outperform medium and large banks. For bank control and bank origin however the story is different. The combined output indicate that private and foreign banks perform better. The multiple output puts private and public banks on equal footing and point to a better performance of domestic over foreign banks. For bank type both output types point to bursary banks as the best performance. They seem to differ significantly in the worst performance however. Credit institutions for the combined output and retail institutions for the multiple output. It should be said however that most of these differences are not statiscally significant. Most pair of confidence intervals will have a non empty intersection as can be seen in Tables A.1 and A.2. This fact is also captured in the nonparametric analysis of covariance shown in Tables A.5, A.6, and A.7. The only significant effects detected are bank origin, marginally, for ϵ^* and bank type for ϕ^* under combined and multiple output, the last result being marginal.

It is important to mention here that in none of the models the variable nonperforming loans (q) seems to affect efficiency (inefficiency) significantly. Berger and Young [21] find mixed evidence regarding the role of nonperforming loans in banking efficiency studies. They find evidence supporting both the bad luck and the bad management hypothesis. The bad luck hypothesis suggests that bank failures are caused primarily by uncontrolled events and thus a better proxy for riskiness of banks could be the concentration of loans and the loans-to-assets ratio. On the other hand, the bad management hypothesis implies that major risks for banking institutions are caused internally, which suggests that supervisors and regulators should analyze bank efficiency along with credit losses and credit risk.

In search for more powerful tests we now investigate the several parametric models discussed in the previous chapter. Tables A.8 to A.10 show goodness of fit statistics for the 14 alternatives implied by the consideration of different hypothesis on the output and different censoring points. The models were fitted using SAS procedures QLIM and NLMIXED. Initial values used for the Tobit alternatives involving the gamma, exponential, and the truncated normal distributions are the estimates of the classical Tobit models. No convergence or singularities were reported by SAS in the fitting process of any of the models. The information measures of Akaike and Schwarz were used to pick the best model for each response. The truncated normal (no Tobit censoring) was the

best fit for DEA residuals. For DEA measurements both with combined and multiple outputs the best alternative is provided by the Tobit censored at 1 defined by the gamma distribution.

Tables A.11, A.12 and A.13 show the results of estimation for the best models. Table A.14 shows the significance of each effect of interest by means of a likelihood ratio test. The models seem to be more informative in regard to technical effects than the ancovas. Significance of effects change with the model used. We see agreement only in bank nature and nonperforming loans. These two effects are not significant in any of the models.

As a further check on model adequacy we use the conditional moment test of specification described in Greene [16]. This is as follows. Let $r(y, x, \theta)$ be a vector of m moment conditions, where y is the response variable, x is the vector of exogenous variables and θ is the unknown parameter, if the model is properly specified, $E(r(y_i, x_i, \theta)) = 0$ for every i. The sample moments are

$$\bar{r}(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} r(y_i, x_i, \hat{\theta})$$

where $\hat{\theta}$ is the maximum likelihood estimator of θ. Let M be the $n \times m$ matrix whose ith row is $r(y_i, x_i, \hat{\theta})$ and G be the $n \times p$ matrix whose ith row is the gradient of the log-likelihood with respect to θ evaluated at $(y_i, x_i, \hat{\theta})$. Let

$$S = \hat{\Sigma} = \frac{1}{n}[M'M - M'G(G'G)^{-1}G'M].$$

If the model is properly specified $n\bar{r}(\hat{\theta})'S^{-1}\bar{r}(\hat{\theta})$ converges in distribution to a chi-square random variable with m degrees of freedom.

We apply the conditional moment test of specification to the model defined by production residuals under the distributional assumption of truncated normal and for the combined and multiple output efficiency measures defined by the Tobit with gamma truncation. We begin the discussion with the production residuals and the truncated normal distribution $N(\mu_i, \sigma^2)$ at $-\mu_i$. Let

$$\lambda_i = \frac{\phi(\mu_i/\sigma)}{\Phi(\mu_i/\sigma)}.$$

The moment conditions we use are :

1. $y_i - \hat{\mu}_i - \hat{\sigma}\hat{\lambda}_i$;
2. $y_i^2 - \hat{\sigma}^2 - \hat{\lambda}\hat{\sigma}\hat{\mu} - \hat{\mu}^2$;
3. $y_i^3 - \hat{\sigma}^3(\sqrt{2\pi}\Phi(\hat{\mu}/\hat{\sigma}))^{-1} h_3(-\hat{\mu}/\hat{\sigma})$;
4. $y_i^4 - \hat{\sigma}^4(\sqrt{2\pi}\Phi(\hat{\mu}/\hat{\sigma}))^{-1} h_4(-\hat{\mu}/\hat{\sigma})$;

where,

$$h_n(y) = \int_y^{+\infty} (x-y)^n \exp(-0.5x^2)\partial x.$$

For the two Tobit gamma models with parameters (p, λ_i), for observations of the response y greater than one, we compute :

1. $y_i - \hat{p}/\hat{\lambda}_i$;
2. $y_i^2 - \hat{p}(\hat{p}+1)/\hat{\lambda}_i^2$;
3. $ln(y_i) - \Psi(\hat{p}) + ln(\hat{\lambda}_i)$;
4. $1/y_i - \hat{\lambda}_i/(\hat{p}-1)$;

where $\Psi(.)$ is the digamma function. For the censored observations we compute :

1. $[(G_{\hat{p}+1}(\hat{\lambda}_i) - G_{\hat{p}}(\hat{\lambda}_i))/G_{\hat{p}}(\hat{\lambda}_i)]\hat{p}/\hat{\lambda}_i$;
2. $[(G_{\hat{p}+2}(\hat{\lambda}_i) - G_{\hat{p}}(\hat{\lambda}_i))/G_{\hat{p}}(\hat{\lambda}_i)]\hat{p}(\hat{p}+1)/\hat{\lambda}_i^2$;
3. $\rho(\hat{p}, \hat{\lambda}_i)/G_{\hat{p}}(\hat{\lambda}_i) - \Psi(\hat{p}) + ln(\hat{\lambda}_i)$;
4. $[(G_{\hat{p}-1}(\hat{\lambda}_i) - G_{\hat{p}}(\hat{\lambda}_i))/G_{\hat{p}}(\hat{\lambda}_i)]\hat{\lambda}_i/(\hat{p}-1)$;

where

$$\rho(p, \lambda) = \frac{1}{G_p(\lambda)\Gamma(p)} \int_0^1 \lambda^p x^{p-1} \exp(-\lambda x) ln(x) \partial x.v$$

The moment conditions for the gamma distribution may be seen in Greene [16].

The chi-square statistics we find for the truncated normal model for the production residuals and the gamma censoring for the single and combined outputs are 0.57, 0.050 and 0.039 clearly non significant.

Which model should we choose ? Our criterion was to pick the model that would mimic the direction of performance of the sample means for all significant effects. The only model showing the proper signs and parameters estimates with this property was the response defined by the multiple output DEA measurement ϕ_2^*. Significant effects indicated by this model are bank type and bank origin. Domestic banks outperform foreign banks and the significance in bank type is due only to pairwise contrasts with the level retail. Finally we mention that these results marginally agree with those provided by the corresponding ancova and both models show approximately the same Pearson correlation between observed and predicted values (about 40%).

4.4 Summary and Conclusion

Output oriented efficiency measurements, calculated under the assumption of variable returns to scale, in the context of Data Envelopment Analysis were investigated for brazilian banks. In this analysis bank outputs investment securities, total loans and demand deposits are analyzed combined in a single measure and as a multiple output

vector to produce different DEA measurements of efficiency based on inputs labor, loanable funds, and stock of physical capital. The intermediation approach is followed and for each measure of efficiency several statistical models are considered as modeling tools to assess the significance of technical effects bank nature, bank type, bank size, bank control, bank origin, and nonperforming loans. The year of analysis is 2001.

The competing statistical models are justified in terms of the stochastic properties of the production responses in the DEA context. The range of model alternatives include the use of nonparametric analysis of covariance, the fit of the truncated normal and the exponential distribution and a general class of Tobit models allowing for heteroscedasticity. All parametric models are fit via maximum likelihood.

The response variable leading to the most informative statistical model uses as response the multiple output-input production model. DEA is oriented to output and computed under the assumption of variable returns to scale. The statistical model chosen is a like a Tobit regression induced by a gamma distribution.

The methodological contributions of the article are as follows. Firstly new alternatives to measure bank output are suggested with the objective of making banks more comparable and to reduce variability and outliers. Secondly it is suggested a collection of statistical models that one could use in a DEA application.

The empirical findings are that domestic banks outperform foreign banks and that all levels of bank type outperform retail with no other pairwise contrasts being significant. None of the models show a significant association of the response with nonperforming loans.

Relevant questions to the administration of the Central Bank of Brazil like the indication of a cut off point for inefficiency measures that would be indicative of bank failure or excessive risk taking, the effect on efficiency of privatization of a public bank or of selling a private bank to a foreign institution as well the effect of merging and acquisitions on bank efficiency were not addressed and cannot be answered in the present study. The reason for this is twofold. Firstly the measure of risk considered in the study, nonperforming loans, is not significant. Secondly to properly address the issues of risk much more complex models are necessary. A panel data structure and past information on other risk and efficiency measures (as cost and revenue efficiencies) will have to be investigated as well.

Chapter 5

Evaluating the Significance of Factors Effects in Output Oriented DEA Measures of Efficiency by a Randomization Process

In this part it is shown that despite of the argument presented by Simar and Wilson [42] against the statistical inference traditionally used in the two-stage approach, the analysis of covariance of non-parametric Data Envelopment Analysis (DEA) estimates ($\hat{\delta}_i = 1/\hat{\phi}_i$) on contextual can be valid. For that, a randomization process is applied to the treatments 10000 times and the resulting p-values of a parametric analysis of covariance with the p-values of the same model, applied to the 'original' data, are compared. The aim is to verify the statistical fundaments on what this kind of analysis could be applied.

5.1 Analysis of Covariance

Similar to equation 4.8, a parametric analysis of covariance is used. But instead of using as responses the rankings r_i of the observations on the variables under investigation, Data Envelopment Analysis (DEA) estimates ($\hat{\delta}_i = 1/\hat{\phi}_i$) for single and multiple output are used. In the general linear model, the basic statistical assumption is that the observed values of the dependent variable can be divided as the sum of two parts : a linear function of the independent coefficients and a random noise component. In other words, the following model is used

$$\hat{\delta}_i = z_{0i}\alpha_0 + \cdots + z_{bi}\alpha_b + u_i \tag{5.1}$$

The u_i are independent $N(0, \sigma^2)$ non observable errors. Under normality, the least-squares estimates are the maximum likelihood estimates. The significance levels and

confidence limit intervals provided by the SAS GLM procedure are based on this assumption but can be good approximations in many other cases.

5.2 Randomization Process

The randomization process is applied with the aim of verifying the validity of the inference of the parametric analysis of covariance for DEA measurements computed for combined and multiple outputs, according to equation (5.1). Randomization avoids the effects of systematic biases that can exist and provides a basis for the assumptions underlying the analysis.

In this case the randomization does not involve blocking and consists on randomly permuting the overall order of the runs, assigning them to the response variable. At the end, each level of a treatment appears once in the completely randomized design.

The variables considered as treatments are bank nature (n), bank type (t) , bank size (s), bank control(c) and bank origin (o). For nonperforming loans (q), the original order was kept.

It is expected, using this process, to obtain similar p-values as those obtained in the parametric analysis of covariance, presented in Tables A.17 and A.19, so inference on them could made. The observations could be considered independent if the subjects are randomly assigned to the treatment levels and if variables associated with the conduct of the experiment are also randomized.

The aleatorization algorithm to be implemented using software SAS [2], consists on :

Loop over the next five steps 10000 times :

1. Generate random numbers from 1 to 94 (number of bank units) ;
2. Identify the number of times each treatment is applied to the bank units, being nt_1 the number of units that received the first treatment, nt_2 the number of units that received the second treatment and so on ;

   A treatment is considered as each different combination of the factors analysed (Bank Nature, Bank Type, Bank Size, Bank Control and Bank Origin) ;
3. Associate to the first nt_1 bank units of the sample randomly generated treatment 1, to the next $nt_1 + 1$ until nt_2 treatment 2, and so on ;

   Keep the original order of the dependent variable (DEA efficiency) and nonperforming loans ;
4. Run a parametric analysis of covariance for DEA measurements computed for a combined and multiple output on nonperforming loans and the randomized treatments ;
5. Store the F values of the model for each variable.

   End Loop.

   At the end of the loop we have a matrix of 10000 F values for 6 variables.

6. Apply a parametric analysis of covariance for DEA measurements computed for a combined output on the original order of treatments and nonperforming loans ;
7. Store the 'original' F value and p-values of the previous model of each variable ;
8. For each variable, calculate the number of F values of the randomized process that exceeds the 'original' F value. Dividing this number by 10000 we obtain a p-value to compare with the 'original' p-value of the 'original' F test.

5.3 Empirical Results

As in Table A.14, the parametric analysis of covariance in Tables A.17 and A.19 shows that the only significant effects detected under combined and multiple outputs are bank type for ϕ^*, the last result being marginal.

In Tables A.18 and A.20, there are the simulation results to evaluate if inference is valid according to the validation process previously explained. P-values of the parametric models for the combined and multiple output are compared with the p-values of the simulation using the randomized process. As expected, the p-values of both cases are really close to each other, specially for the categorized variables. For the variable nonperforming loans, the same level of proximity was not obtained, but still the conclusion of the parametric models remain correct.

Chapter 6

Bootstrap Procedures

6.1 Simple Bootstrap Bias Corrected Confidence Intervals for Factors Effects of Brazilian Banks DEA Efficiency Measures

The use of bootstrap methods is attractive when distributional properties of an estimator are unknown and the respective standard error is not easily obtained. This method is used to check the results of the best model adjusted in chapter 4.2, to verify their accordance in the sense of parameter's significance, their bias and also to calculate DEA efficiency confidence intervals. In the first part of this chapter it is briefly explained the bootstrap method applied to the data set, according to Souza [8].

If the distribution of the parameters has appreciable bias, the performance of the basic percentile confidence interval is affected. To avoid this problem, a bias corrected percentile confidence interval was considered. How to calculate the bias, its significant test and the confidence interval is included in subsection (6.1.1).

6.1.1 The Bootstrap Algorithm

The bootstrap applied considers the regression model defined in chapter 4.1, equation (4.1) :

$$y_i = g(x_i) - \epsilon_i \tag{6.1}$$

where the inefficiency errors were supposed to be generated from a truncated normal distribution. The subsequent algorithm reproduces the general steps to be followed in a SAS program to implement the bootstrap theory. The bootstrap sample size is 1500. By Hall [6], 1000 replications should be enough.

Bootstrap Algorithm

1. Calculate the estimated DEA measures from Definition 1 (ϕ_i^*) for each production unit, using the inputs and outputs of the sample data ;
2. Get the inefficiency errors ϵ_i^* from equation (4.4) ;
3. Obtain the maximum likelihood estimates of the parameters $\hat{\beta}$ in (4.5), considering the inefficiencies have a left-truncated normal distribution, with mean ($\hat{m}_i$) and variance ($\hat{v}_i$), according to equations (4.6) and (4.7) ;
4. Loop over the next steps 1500 times to obtain the bootstrap sample of the parameters $\hat{\beta}$ and efficiency (ϕ_i^*) estimates, for each unit $i = 1, \cdots, n$:

   4.1 Generate errors (ϵ_i^b) from a left-truncated normal distribution with mean ($\hat{m}_i$) and variance ($\hat{v}_i$) ;

   4.2 Adjust the regression model as in item 3, supposing the truncated normal distribution, but using the bootstrap inefficiencies (ϵ_i^b) and obtain the bootstrap parameter estimates ($\hat{\beta}^b$) ;

   4.3 Compute bootstrap efficiency measures $\hat{\phi}_i^b = 1 + \frac{\epsilon_i^b}{y_i}$;

   4.4 Keep the bootstrap estimated parameters ($\hat{\beta}^b$) obtained in subitem 4.2, as also the bootstrap efficiency measures ($\hat{\phi}_i^b$) from subitem 4.3 ;
5. End Loop.
6. For the parameters and DEA efficiencies, based on the bootstrap sample :

   6.1 Calculate the estimated bias and test its significance ;

   6.2 Construct the bias corrected percentile confidence intervals.

Bias and Significance Test The relative bias of the parameters can be obtained by :

$$100\frac{\bar{\hat{\beta}}^b - \hat{\beta}}{\hat{\beta}} \tag{6.2}$$

where $\bar{\hat{\beta}}^b$ is the bootstrap mean.

The significance of the bias can be tested knowing that

$$z = \sqrt{B}\frac{\bar{\hat{\beta}}^b - \hat{\beta}}{\sqrt{Var(\hat{\beta})}} \;\; is \;\; N(0,1) \tag{6.3}$$

under the null hypothesis of bias inexistency, where B is the bootstrap sample size.

Confidence Intervals The bias corrected percentile confidence interval for a given parameter β, at $100(1-\alpha)\%$ significance level is given by :

$$[H^{-1}(\Phi(2z_0 - z_{\alpha/2})), H^{-1}(\Phi(2z_0 + z_{\alpha/2}))] \tag{6.4}$$

where $z_0 = \Phi^{-1}(H(\hat{\beta}))$, $\Phi(x)$ is the standard normal distribution function and $H(u)$ is the bootstrap distribution function of $\hat{\beta}^b$.

The same idea of subsection (6.1.1) can be used for the DEA efficiencies.

6.1.2 Bootstrap Results

In this part bootstrap results are showed. Based on the descriptive statistics and the Kolmogorov-Smirnov test for normality in Table A.21, only five distributions of the parameters follow normality : credit and business type (t_1 and t_2), large and medium size (s_1 and s_2) and bank origin (o_1) marginally. Also from Table A.21 it is observed that the relative bias is extremely high for the intercept, commercial and multiple nature (n_1 and n_2) variables, as also for the variance. The first three distributions are the most assimetrics. The bias is significantly different from zero to all parameters as reported by the z values.

Table A.22 provides the bootstrap confidence intervals and means, estimated confidence intervals and parameters from the truncated normal model, bias and bias corrected parameters. Differently from the results of the model adjustment, it can be seen that for the bootstrap the parameters for large and medium size (s_1 and s_2) and private control c_1 are not significantly different from zero, but for the truncated normal model, their significance was marginal (Table A.11). The bias corrected parameters have the same sign as the original parameters of the model even for the most assimetric distributions ($intercept$, n_1 and n_2).

6.2 Estimation and Inference in a Double Bootstrap Applied to the DEA Efficiency Measures of Brazilian Banks

The objective of this part is to compute efficiency measures for commercial banks in Brazil and to verify the influence of contextual variables on it. A double bootstrap proposed by Simar and Wilson [42] is applied. Initially, a DEA is performed as in chapter 4.1. But in the second stage, instead of only regressing the contextual variable on the resulting efficiency, a double bootstrap procedure is applied, allowing for inference in the regression model, according to Simar and Wilson [42], since it corrects for the correlation and bias problem. These problems are consequences of the lack of specification of the data generating process (DGP) of the DEA measures.

We apply the double bootstrap for the best parametric model for DEA measurements from combined output ($A.12$). It is based on a Tobit model with censoring at 1 and residuals with gamma distribution with shape parameter P.

6.2.1 DEA Efficiency Measures

For the construction of efficiency measures, a production frontier can be defined in different ways. Basically, the most used efficiency measures are based on Data Envelopment Analysis and on the Free Disposal Hull estimates of the production set (Ψ).

Kneip *et al.* [32] and Park *et al.* [38] describe the tools for inference analysis based on asymptotic or bootstrap results and Simar and Wilson [40] provide a survey of the inference results for DEA/FDH efficiency scores.

As in chapter 4.1, define a production process with n production units that needs a set of p inputs x to produce a quantity y of s outputs and denote by $Y = (y_1, ..., y_n)$ the $s \times n$ output matrix and by $X = (x_1, ..., x_n)$ the $p \times n$ input matrix. The Farrell-Debreu efficiency measure is defined as :

$$\phi(x, y) = sup\left\{\phi | (x, \phi y) \in \Psi\right\}. \tag{6.5}$$

The Farrell-Debreu measure $\phi(x, y) \geq 1$ and the excess over 1 means the percentage of output that could be increased to achieve efficiency, given that the input/output set of the firm is (x, y).

The general formulation for a production set is given by :

$$\Psi = \left\{(x, y) \in \Re_+^{p+s} | x\, produces\, y\right\}. \tag{6.6}$$

But the production set is not observable and must be estimated. And as previously explained, DEA and FDH production sets are the two most used options to be plugged in the input/output oriented measure, and are explained below.

For the DEA output oriented efficiency measure, the production set to be plugged in is the smallest free disposal convex set that contains the input/output data set, and is given by :

$$\hat{\Psi}_{DEA} = \left\{(x, y) \in \Re_+^{p+s} | y \leq \sum_{i=1}^{n} \gamma_i y_i;\ x \geq \sum_{i=1}^{n} \gamma_i x_i\, for\, (\gamma_1, \cdots, \gamma_n)\right\} \tag{6.7}$$

$$s.t. \sum_{i=1}^{n} \gamma_i = 1;\ \gamma_i \geq 0,\ i = 1, \cdots, n. \tag{6.8}$$

Under free disposability, if $(x, y) \in \Psi$, then $(x', y') \in \Psi$, as soon as $x' \geq x$ and $y' \leq y$, and without the assumption of convexity, from Deprins *et al.* [28], we have the Free Disposal Hull production set :

$$\hat{\Psi}_{FDH} = \left\{(x, y) \in \Re_+^{p+s} | y \leq y_i;\ x \geq x_i,\ i = 1, \cdots, n\right\}. \tag{6.9}$$

The construction of the DEA frontier is based on linear programming methods and the efficiency measures are subsequently calculated relative to this surface. Charnes *et al.* [24] published the first paper using the DEA term, proposing a model for the input orientation and constant returns to scale. Previous work on that as Boles [3] and Afriat [4] did not receive attention. Banker *et al.* [20] proposed an extension for the estimation of DEA efficiencies considering variable returns to scale (VRS).

For the output case (VRS) the linear programming problem is given by **Definition 1** in chapter 4.1. The solution for ϕ in the linear program for a given point (x, y) is the estimated DEA efficiency measure $(\hat{\phi}_{DEA}(x, y))$ of the unit. In this chapter the index DEA is used to differentiate from the FDH efficiency measure.

The difference between constant returns to scale and variable returns to scale is given by the addition of the convexity constraint : $n1'\lambda = 1$, which generates a tighter production frontier that envelopes the whole data set. The linear programming problem must be solved for each financial institution. Those with efficiency values equal to 1 are on the frontier, it means, they achieved maximum efficiency.

The nonparametric FDH efficiency measure estimates $(\hat{\phi}_{FDH}(x, y))$ for the output oriented case are obtained substituting Ψ by $\hat{\Psi}_{FDH}$ in equation (6.5), and for a given point (x, y) it results in :

$$\hat{\phi}_{FDH}(x, y) = sup\left\{\phi|(x, \phi y) \in \hat{\Psi}_{FDH}\right\}. \tag{6.10}$$

From Simar and Wilson [41] it can be computed by :

$$\hat{\phi}_{FDH}(x, y) = \max_{i \in D(x,y)} \left\{ \min_{j=1,\ldots,p} (\frac{y_i^j}{y^j}) \right\} \tag{6.11}$$

where for a vector a, a^j denotes the jth element of a, and

$$D(x, y) = \{i|(x_i, y_i) \in \chi_n,\ x_i \leq x, y_i \geq y\}, \tag{6.12}$$

$\chi_n = \{(x_i, y_i), i = 1, \cdots, n\}$ and $D(x, y)$ is the set of sample points dominating the point of interest (x, y).

In Simar and Wilson [41] can be found a summary of asymptotic properties of $\hat{\Psi}_{DEA}$ and $\hat{\Psi}_{FDH}$. For example, from Korostelev *et al.* ([33] and [34]), for $p = 1$ and $s \geq 1$, under free disposability we have :

$$d_\Delta(\hat{\Psi}_{FDH}, \Psi) = O_p(n^{-\frac{1}{s+1}})$$

and under free disposability and convexity of the production set :

$$d_\Delta(\hat{\Psi}_{DEA}, \Psi) = O_p(n^{-\frac{2}{s+2}})$$

where $d_\Delta(\cdot,\cdot)$ is the Lebesgue measure (giving the volume) of the difference between the two sets.

When s is small, the rates of convergence are larger, indicating the superiority of the $\hat{\Psi}_{DEA}$ estimator in this aspect. But convergence is obtained only if the DEA estimator is consistent. And for consistency, the convexity assumption must hold what is not necessary in the FDH context.

For the more general case when $p \geq 1$, from Park *et al.* [38] and Kneip *et al.* [32] we have the following FDH and DEA efficiency measures results :

$$\hat{\phi}_{FDH} - \phi = O_p(n^{-\frac{1}{p+s}})$$

and under free disposability and convexity of the production set :

$$\hat{\phi}_{DEA} - \phi = O_p(n^{-\frac{2}{p+s+1}})$$

where $\hat{\phi}_{FDH}$ and $\hat{\phi}_{DEA}$ are the FDH and DEA efficiency measure estimators respectively.

As it can be observed, the rates of convergence depend on the size of the output and input vector (p and s), the greater they are, the slower are the rates of convergence. It is known as the 'curse of dimensionality'.

Again, superiority (slightly faster convergence rate) of the $\hat{\Psi}_{DEA}$ estimator over $\hat{\Psi}_{FDH}$ exists in case of convexity of Ψ. This result is a consequence of the fact that $\hat{\Psi}_{FDH} \subseteq \hat{\Psi}_{DEA} \subseteq \Psi$, it means, both estimators are biased by construction. In the output orientation, this relationship implies that :

$$\hat{\phi}_{FDH}(x,y) \leq \hat{\phi}_{DEA}(x,y) \leq \phi(x,y). \tag{6.13}$$

6.2.2 Double Bootstrap in a Two-stage Approach

In their paper, Simar and Wilson show that the statistical inference traditionally used in two-stage approaches, the regression of non-parametric Data Envelopment Analysis (DEA) estimates $\hat{\phi}_i$ on contextual variables, are invalid. The usual parametrization is

$$\phi_i = z_i\beta + \epsilon_i \geq 1 \tag{6.14}$$

where z_i is the $i-th$ observation of the $Z \in \Re^r$ vector.

But since ϕ_i is not observable it must be estimated, and in equation (6.14) it can be substituted by $\hat{\phi}_i$, the nonparametric DEA estimate, and the following model could be estimated :

$$\hat{\phi}_i = z_i\beta + \xi_i \geq 1. \tag{6.15}$$

As demonstrated by Simar and Wilson [42], in this case standard inference is flawed due to the following reasons :

– although consistent, $\hat{\phi}_i$ has a strictly negative bias in finite samples ;
– the error term ϵ_i is correlated and is also correlated with the contextual variables z_i.

Since $\hat{\phi}_i$ is negatively biased in finite samples, it should be corrected for the bias, resulting in the bootstrap bias-corrected estimator of ϕ_i :

$$\hat{\hat{\phi}}_i = \hat{\phi}_i - \hat{BIAS}(\hat{\phi}_i) \tag{6.16}$$

The explanation comes from the fact that :

$$\hat{\phi}_i = E(\hat{\phi}_i) + u_i \tag{6.17}$$

where $E(u_i) = 0$. Besides, by definition, the bias of $\hat{\phi}_i$ is :

$$BIAS(\hat{\phi}_i) \equiv E(\hat{\phi}_i) - \phi_i. \tag{6.18}$$

Substituting $E(\hat{\phi}_i)$ from (6.17) in (6.18) we get :

$$\phi_i = \hat{\phi}_i - BIAS(\hat{\phi}_i) - u_i. \tag{6.19}$$

Finally, substituting ϕ_i in (6.14) results in :

$$\hat{\phi}_i - BIAS(\hat{\phi}_i) - u_i = z_i\beta + \epsilon_i \geq 1 \tag{6.20}$$

what justifies the regression in (6.15), considering that, asymptotically, $u_i \to 0$ and also $BIAS(\hat{\phi}_i)$ and consequently, $\hat{\phi}_i$ is consistent.

Since $BIAS(\hat{\phi}_i)$ does not has zero mean and can be estimated by bootstrap methods, differently from u_i that has zero mean and cannot be estimated, the regression to be estimated becomes :

$$\hat{\hat{\phi}}_i \approx z_i\beta + \xi_i \geq 1, \tag{6.21}$$

on which maximum likelihood estimation can be applied, providing consistent estimates.

Some assumptions for the model are required, as explained in Simar and Wilson [42], which are reproduced below :

– **Assumption A1** : the sample observations (x_i, y_i, z_i) in $\gamma_n = \{(x_i, y_i, z_i)\}_{i=1}^n$ are realizations of identically, independently distributed random variables with probability density function $f(x, y, z)$ which has support over $\Psi \times \Re^r$, where $\Psi \subset \Re_+^{p+s}$ is a production set defined by

$$\Psi = \left\{(x, y) \in \Re_+^{p+s} | x \, produces \, y\right\}. \tag{6.22}$$

– **Assumption A2** : the conditioning in $f(\phi_i|z_i)$ in the joint density $f(x_i, \eta_i, \phi_i, z_i) = f(x_i, \eta_i|\phi_i, z_i)\ f(\phi_i|z_i)\ f(z_i)$ operates through the following mechanism :

$$\phi_i = \psi(z_i, \beta) + \epsilon_i \geq 1, \tag{6.23}$$

where ψ is a smooth, continuous function, β is a vector of (possibly infinitely many) parameters, ϵ_i is a continuous iid random variable, independent of z_i and $\eta_i = [\eta_{i1}\ \ \eta_{i2}\ \ \cdots\ \ \eta_{i,s-1}]$,

$$\eta_{ij} = \begin{cases} arctan(\frac{y_{i,j+1}}{y_{i1}}), & if\ \ y_{i1} > 0 \\ \frac{\pi}{2}, & if\ \ y_{i1} = 0, \end{cases}$$

for $j = 1, \cdots, s-1$ and $y_i = [y_{i1} \cdots y_{is}]$.

– **Assumption A3** : ϵ_i in (6.22) is distributed $N(0, \sigma_\epsilon^2)$ with left-truncation at $1 - \psi(z_i, \beta)$ for each i.

– **Assumption A4** : Ψ is closed and convex; $y(x)$ is closed, convex and bounded for all $x \in \Re_+^p$; and $\chi(y)$ is closed and convex for all $y \in \Re_+^s$, where $y(x) \equiv \{y|(x, y) \in \Psi\}$ and $\chi(y) \equiv \{x|(x, y) \in \Psi\}$ are the sections of the production set Ψ.

– **Assumption A5** : $(x, y) \notin \Psi$ if $x = 0$, $y \geq 0$, $y \neq 0$, i.e., all production requires use of some inputs.

– **Assumption A6** : for $x' \geq x$, $y' \leq y$, if $(x, y) \in \Psi$ then $(x', y) \in \Psi$ and $(x, y') \in \Psi$, i.e., both inputs and outputs are strongly disposable.

– **Assumption A7** : for all $(x, y) \in \Psi$ such that $(\phi^{-1}x, y) \notin \Psi$ and $(x, \phi y) \notin \Psi$ for $\phi > 1$, $f(x, y|z)$ is strictly positive, and $f(x, y|z)$ is continuous in any direction toward the interior of Ψ for all z.

– **Assumption A8** : for all (x, y) in the interior of Ψ, $\phi(x, y|\Psi)$ is differentiable in both its arguments. Where

$$\phi(x, y|\Psi) = \frac{w(\phi(x, y|\Psi)y)}{w(y)}, \tag{6.24}$$

and $w(y) = y'y$.

In summary, assumption A1 represents the separability condition between the input x output space and the space of values of z and A2 states how z influences the efficiencies. Assumption A3 associates the truncated normal distribution for the distribution of the error term ϵ_i, A4 is related to some classical mathematical constraints in standard microeconomic theory of the firm. The inexistence of *free lunch* is characterized by A5 and free disposability of inputs and outputs by A6. To assure consistency of the estimates of Ψ and ϕ_i assumptions A7 and A8 are required. The main reason of defining these assumptions is to specify a semi-parametric data generating process for the

vector (x_i, y_i, z_i). But the problem of correctly estimating ϕ_i and the parameters of the regression still remains.

The correlation in the error term in equation (6.15) comes from the fact that each estimated efficiency measure $\hat{\phi}_i$ is calculated using all the observations (x_i, y_i) in $\gamma_n = \{(x_i, y_i, z_i)\}_{i=1}^n$ through the estimator of the production set $(\hat{\Psi})$. It means, if the value of one observation changes, the estimated frontier will be affected and consequently, some (or all) efficiency estimates.

The correlation of the error term with the dependent variable is a consequence of A2. From Assumption A1 we have that the observations of γ_n are independently drawn, although from A2 we have that x_i and y_i are correlated with z_i. This assumption assures that the conditional relation between ϕ_i and z_i is given by :

$$\phi_i = \varphi(z_i, \beta) + \epsilon_i \geq 1 \tag{6.25}$$

where φ must be a smooth, continuous function, β is the parameter's vector and ϵ_i are independent identically distributed, also independent of z_i.

More details can be obtained in Simar and Wilson [41].

Asymptotically, the bias and correlation problems disappear at a slow rate, it assures consistency of β and σ. But in the case of finite samples, it is necessary to correct for these problems so as to be able to make inference about β. The authors suggest two bootstrap procedures. The first one permits inference but does not correct for the bias. The second one corrects for both problems : bias and correlation.

For both alternatives (presented below) and the simple regression, inference performances in the second-stage approach were checked by Monte Carlo experiments. Simar and Wilson consider the coverage of confidence intervals and the root mean square error (RMSE) of the coefficients to evaluate the bootstraps. In general, coverages improve as n increases and become worst as $p + q$ increases, as it reduces the precision of the estimates in the second stage. Comparing algorithms #1 and #2, the second reveals improved coverages in number of cases, but as in algorithm #1, the coverages obtained with the simple regression are broadly similar to those in algorithm #2. Considering the RMSE, for $p = q = 1$, 2 or 3, and sample size of 100, better results are obtained for the simple regression. By the other side, when n increases to 400, with $p = q = 1$ or 2, algorithm #2 provides lower RMSE for the intercept and slope estimators of the efficiency.

Algorithm 1 This bootstrap is built to improve inference, the double bootstrap takes also the bias into account.

1. Calculate the DEA efficiency measure $\hat{\phi}_i$, for $i = 1, \cdots, n$;

2. Based on a gamma regression (Tobit truncated at 1) with shape parameter p of $\hat{\phi}_i$ on z_i, estimate $\hat{\beta}$ and $\hat{p}$ by MLE, deleting the spurious $\hat{\phi}_i = 1$, using $m < n$ observations ;
3. For $b_1 = 1, \cdots, L_1$ obtain the bootstrap estimates $\hat{\beta}^*$ and the shape parameter $\hat{p}^*$ (for $i = 1, \cdots, m$), based on the following steps :
   3.1. Generate ϕ^*_{iG}, where ϕ^*_{iG} is $G(\hat{p}, \hat{\lambda}_i)$, where $\hat{\lambda}_i = \exp(-z_i\hat{\beta})$;
   3.2. Compute ϕ^*_i. If $\phi^*_{iG} \leq 1$ then $\phi^*_i = 1$, else $\phi^*_i = \phi^*_{iG}$;
   3.3. Based on the gamma regression (Tobit truncated at 1) of ϕ^*_i on z_i, estimate $\hat{\beta}^*$ and the shape parameter $\hat{p}^*$ by MLE ;
4. Construct confidence intervals based on $\hat{\beta}$ and $\hat{p}$ and the bootstrap estimates $\hat{\beta}^*$ and $\hat{p}^*$ for the β vector and p shape parameter.

Algorithm 2 - Double Bootstrap The double bootstrap procedure, suggested by Simar and Wilson [42], provides ways of constructing confidence intervals for the second stage regression that allow for valid inference on the parameters of the model. It can be implemented following the steps described below :

1. Calculate the DEA efficiency measure $\hat{\phi}_i$, for $i = 1, \cdots, n$;
2. Based on a gamma regression (Tobit truncated at 1) with shape parameter p of $\hat{\phi}_i$ on z_i, estimate $\hat{\beta}$ and $\hat{p}$ by MLE, deleting the spurious $\hat{\phi}_i = 1$;
3. For $b_1 = 1, \cdots, L_1$ obtain the bootstrap estimates $\hat{\phi}^*_{ib}$ (for $i = 1, \cdots, n$), based on the following steps :
   3.1. Generate ϕ^*_{iG}, where ϕ^*_{iG} is $G(\hat{p}, \hat{\lambda}_i)$, where $\hat{\lambda}_i = \exp(-z_i\hat{\beta})$;
   3.2. Compute ϕ^*_i. If $\phi^*_{iG} \leq 1$ then $\phi^*_i = 1$, else $\phi^*_i = \phi^*_{iG}$;
   3.3. Define $x^*_i = x_i$ and $y^*_i = \frac{y_i\hat{\phi}_i}{\phi^*_i}$;
   3.4. Redefine a new production set $\hat{\Psi}^*$ based on $Y^* = [y^*_1 \ \ldots \ y^*_n]$ and $X^* = [x^*_1 \ \ldots \ x^*_n]$ and calculate $\hat{\phi^*_{ib}} = \phi(x_i, y_i|\hat{\Psi}^*)$;
4. For each observation, calculate the bias-corrected estimator $\hat{\hat{\phi}}_i = 2\hat{\phi}_i - \sum_{b=1}^{L_1} \frac{\hat{\phi}^*_{ib}}{L_1}$;
5. Based on a gamma regression (Tobit truncated at 1) with shape parameter p of $\hat{\hat{\phi}}_i$ on z_i, estimate $\hat{\hat{\beta}}$ and $\hat{\hat{p}}$ by MLE ;
6. For $b_2 = 1, \cdots, L_2$ calculate the bootstrap estimates $\hat{\hat{\beta}}^*$ and $\hat{\hat{p}}^*$, based on the following steps :
   6.1. Generate ϕ^*_{iG}, where ϕ^*_{iG} is $G(\hat{\hat{p}}, \hat{\hat{\lambda}}_i)$, where $\hat{\hat{\lambda}}_i = \exp(-z_i\hat{\hat{\beta}})$;
   6.2. Compute ϕ^{**}_i. If $\phi^*_{iG} \leq 1$ then $\phi^{**}_i = 1$, else $\phi^{**}_i = \phi^*_{iG}$;
   6.3. Based on a gamma regression (Tobit truncated at 1) with shape parameter p of ϕ^{**}_i on z_i, estimate $\hat{\hat{\beta}}^*$ and $\hat{\hat{p}}^*$ by MLE ;
7. Construct confidence intervals based on the bootstrap estimates $\hat{\hat{\beta}}^{**}$ and $\hat{\hat{p}}^{**}$ for the β vector and p shape parameter.

To compute the bias-corrected estimates $\hat{\hat{\phi}}_i$ the number of replications L_1 suggested by the authors is 100. The small number of replications is justified because only the mean of the generated parameter is obtained from the algorithm. In the second bootstrap, the number of replications must be much greater, 1000 by Hall [6], since the objective is to obtain confidence intervals for the parameters. Simar and Wilson [42] used 2000 replications for the second loop where the truncated regression model is bootstrapped.

Empirical Results

The bootstrap is implemented for the parametric model in Table $A.12$ for combined output, a Tobit with censoring at 1, and gamma distribution with shape parameter P.

In the second step of algorithms 1 and 2, Simar and Wilson suggest to run the parametric regression excluding the observations whose estimated efficiencies equal 1 (Table $A.23$) and to consider these estimates on the bootstrap. They argue that the probability mass at 1 is an artifact of finite samples and is not related to the *true* model specified in 6.23. From Table $A.23$ some estimated parameters change significantly when compared to those specified in Table $A.12$. The strongest difference is for the shape parameter P, the 'new' confidence interval $(3.28, 5.99)$ not even includes the original estimated parameter (3.08). Also, the variable s_2 (medium size) is considered significant when excluding estimated efficiencies equal to 1, differently from before.

Algorithm 1 This algorithm does not include a bias correction, but it is applied for the bootstrap mean, so as to calculate a bias corrected percentile confidence interval, according to 6.4. The results are presented in Table $A.24$. In step 3 the loop has 2500 replications.

Without $\hat{\phi}^{*1} = 1$ in step 2 the conclusions differ from $A.12$. The shape parameter is not included in the confidence interval and the variable s_2 (medium size) is considered significant.

The Pearson correlation between observed and predicted values is 60%.

Algorithm 2 The results are related to a double bootstrap applied for the best parametric model for DEA measurements from combined output ($A.12$). It is based on a Tobit model with censoring at 1 and residuals with gamma distribution with shape parameter P. The first loop (step 3) has 1000 replications and the second one (step 6) 2000. We observe that observation 80 has an extremely low value for the estimated efficiency, but since it was not influent on the parameter estimates, we decided to keep it on the analysis.

In Table $A.25$ we have the double bootstrap means and confidence intervals for the Tobit model censored at 1, gamma distribution with shape parameter P (without $\hat{\phi}^{*1} = 1$ in step 2) and respective measures of the original model, as in Table $A.12$.

The original shape parameter (3.08) is not included in the double bootstrap confidence interval $(3.16, 5.78)$. Excluding $\hat{\phi}^{*1} = 1$ in the second step of the bootstrap we remark that variable $s2$ (medium size) is considered significant, but marginally.

The Pearson correlation between observed and predicted values is 60%.

Algorithms 1 and 2 are consistent with each other but differ in a similar way to the Tobit model (Table $A.12$). So as to be able to compare the Pearson correlation between observed and predicted values from the original model, we applied a bias corrected bootstrap to it, and the Pearson correlation was also around 60%. It is worthy to remark that also the sigificance of the parameters did not change in this bootstrap, when compared to the model in Table $A.12$.

Chapter 7

A Probabilistic Approach for Brazilian Banks Contextual Variables in Nonparametric Frontier Models

In this chapter we present a probabilistic interpretation of the Farrell-Debreu efficiency scores. The formulation proposed by Daraio and Simar [27] is for a nonparametric frontier model that can also consider external contextual factors (neither outputs nor inputs) that might influence the production process. For that, a probabilistic model is necessary to define the data generating process.

In this context, the new concept of conditional efficiency measure and respective nonparametric estimators are also presented. The previous ideas were developed initially by Cazals *et al.* [23]. The authors also proposed the order-m methods due to the sensitivity of the DEA and FDH to outliers. An empirical evidence of this problem can be found in Wheelock and Wilson [45] in their study of efficiency and technical change in U.S. commercial banking. Basically, the results are more robust since the frontier does not envelope all the data, since they are not constructed using all the observations available, but a subset of it. Instead of this method, we opt for excluding outliers before calculating the efficiency measures.

As pointed out by Daraio and Simar [27], one main difference between the two stage approach and the probabilistic formulation is that the first depends on the separability condition between the contextual variable Z and the input x output set (X, Y), what is not necessary in the second one. This condition implies that the production frontier does not change with a different set of the contextual variable, since it does not depend on that.

Besides, in the two stage approach the Data Envelopment Analysis that incorporates convexity assumption is used, while in the probabilistic formulation, the use of the Free Disposal Hull (FDH) efficiency scores does not require this hypothesis. Another difference is that the probabilistic approach is non parametric while in the two stage we need to specify a parametric function to be able to regress the estimated efficiency

on the contextual variables. In most of the studies, the error term is supposed to follow a truncated normal distribution. Although other authors as Banker [7] and Souza [11] have already studied other possibilities as the use of an exponential distribution.

To verify the influence of the contextual variable z on the production process the FDH efficiency scores conditional and non-conditional on this variable are compared.

In the stochastic approach, the stochastic part of the DGP specified in Assumptions A1 until A8, through the probability density function $f(x, y)$ or the corresponding distribution function $F(x, y)$ is substituted by the following probability function (Simar and Wilson [41]) :

$$H_{XY}(y, x) = P(Y \geq y, X \leq x). \tag{7.1}$$

The authors provide the following interpretations and properties :
– "$H_{XY}(y, x)$ 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).
– $H_{XY}(y, x)$ is monotone, non-decreasing in x and monotone non-increasing in y.
– The support of the distribution function $H_{XY}(.,.)$ is the attainable set Ψ ; i.e.,

$$H_{XY}(y, x) = 0 \ \ \forall \ \ (x, y) \notin \Psi.\text{"} \tag{7.2}$$

Applying Bayes' rule in the probability function $H_{XY}(y, x)$ we get :

$$H_{XY}(y, x) = P(X \leq x | Y \geq y)P(Y \geq y) = F_{X/Y}(x|y)S_Y(y) \tag{7.3}$$

and

$$H_{XY}(y, x) = P(Y \geq y | X \leq x)P(X \leq x) = S_{Y/X}(y|x)F_X(x). \tag{7.4}$$

New concepts of efficiency measures can be defined for the input-oriented case and output-oriented case, assuming $S_Y(y) > 0$ and $F_X(x) > 0$:

$$\theta(x, y) = inf\left\{\theta | F_{X/Y}(\theta x|y) > 0\right\} = inf\left\{\theta | H_{XY}(\theta x, y) > 0\right\} \tag{7.5}$$

and

$$\lambda(x, y) = sup\left\{\lambda | S_{Y/X}(\lambda y|x) > 0\right\} = sup\left\{\lambda | H_{XY}(\lambda y, x) > 0\right\}, \tag{7.6}$$

since the support of the joint distribution is the attainable set, boundaries of Ψ can be defined in terms of the conditional distributions.

Comparing with the DEA measures, there is also a difference in the interpretation of the efficiency scores in (7.5) and (7.6) :
– Input case : 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 ;

– Output case : is the proportionate increase in outputs required for the same unit to have zero probability of being dominated, holding input levels fixed.

Considering that the output orientation is of interest for the empirical work, also in this chapter, only this case will be presented in more details. This part of the analysis will be based in the following two output oriented efficiency measures : Free Disposal Hull (FDH) and conditional FDH efficiency measure.

7.1 Unconditional Probabilistic Formulation

The Farrell-Debreu output efficiency measure for a given level of input (x) and output (y) is defined as in equation (6.5) and in the free disposability context is given by

$$\lambda(x,y) = sup\left\{\lambda | S_{Y|X}(\lambda y, x) > 0\right\} \tag{7.7}$$

where $S_{Y|X}(y|x) = P(Y \geq y | X \leq x)$.

And it can be non parametrically estimated by

$$\hat{\lambda}_n(x,y) = sup\left\{\lambda | \hat{S}_{Y|X,n}(\lambda y|x) > 0\right\} \tag{7.8}$$

where $\hat{S}_{Y|X,n}(y|x) = \frac{\sum_{i=1}^{n} I(x_i \leq x, y_i \geq y)}{\sum_{i=1}^{n} I(x_i \leq x)}$.

In practice, it is estimated by

$$\hat{\lambda}_n(x,y) = sup\left\{\lambda | (x, \lambda y) \in \hat{\Psi}_{FDH}\right\} = \max_{i|x_i \leq x}\left\{\min_{j=1,\ldots,q}(\frac{y_i^j}{y^j})\right\} \tag{7.9}$$

because, as observed by Cazals *et al.* [23], it coincides with the FDH estimator.

As already mentioned, the estimated FDH production set is very sensitive to outliers, and consequently, are the estimated efficiency scores. Daraio and Simar [27] 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.

The full frontier gives the full maximum achievable level of output over all production plans that are technically feasible. An alternative benchmark is obtained by defining the expected maximum output achieved by m firms chosen randomly from the population and using a maximum quantity of inputs at level x. In summary, the order-m frontier provides a less extreme frontier in case of outliers. As m increases, the order-m frontier converges to the full frontier. This method was substituted by excluding the outliers before calculating the efficiency measures.

7.2 Conditional Probabilistic Formulation

Cazals *et al.* [23] proposed the use of probabilistic non-parametric frontier models for the univariate case, permitting one input in the input oriented case and one output for the output oriented case. Supposing that the separability condition is not valid and the production frontier is influenced by the contextual variables, they also suggested the introduction of $Z \in \Re^r$ by conditioning the production process on it.

Daraio and Simar [27] extended their approach to the multivariate case. In their paper, the authors explicit the input oriented framework that we adapt here for the output oriented case. Conditioning on $Z = z$, the efficiency measure is given by :

$$\lambda(x, y|z) = sup\left\{\lambda | F_Y(\lambda y|x, z) > 0\right\}, \tag{7.10}$$

where $F_Y(y|x, z) = Prob(Y \geq y|X \leq x, Z = z)$.

As $F_Y(y|x, z)$ is not observable, it is necessary to define a non-parametric estimator for it applying smoothing techniques on z due to the continuity of this variable. Considering the sample size n, the following kernel estimator for $F_Y(y|x, z)$ is defined as :

$$\hat{F}_{Y,n}(y|x, z) = \frac{\sum_{i=1}^{n} I(x_i \leq x, y_i \geq y) K(\frac{z-z_i}{h_n})}{\sum_{i=1}^{n} I(x_i \leq x) K(\frac{z-z_i}{h_n})}, \tag{7.11}$$

where $K(.)$ is the kernel and h_n is the bandwidth. The bandwidth selection suggested by Daraio and Simar [27], the likelihood cross validation criterion, using a k-NN method, is described in Silverman [5].

The smoothing technique is necessary if Z is a continuous variable. The basic idea is of smoothing the conditional distribution function ($\hat{F}_{Y,n}(y|x, z)$) estimation, selecting a bandwidth h which could optimize 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. The choice of the k-NN method results in the choice of a local bandwidth h_{z_i}, always with the same number of observations, it means, k points z_j verifying $|z_j - z_i| \leq h_{z_i}$.

The cross validation criteria evaluates the leave-one-out kernel density estimate of Z, $\hat{f}_k^{(-i)}(Z_i)$, $i = 1, \cdots, n$, for some values of k and choose the one that maximizes the score function :

$$CV(k) = n^{-1} \sum_{i=1}^{n} log(\hat{f}_k^{(-i)}(Z_i)),$$

$$\hat{f}_k^{(-i)}(Z_i) = \frac{1}{(n-1)h_{Z_i}} \sum_{j=1, j\neq i}^{n} K\left(\frac{Z_j - Z_i}{h_{Z_i}}\right).$$

Also from Silverman [5], for a specific kernel function, the discrepancy between the density estimator and the true density $f(x)$ can be measured by the mean integrated square error (MISE) :

$$MISE(h) = \int_x \left\{ E(\hat{f}_h(x) - f(x)) \right\}^2 dx + \int_x VAR(\hat{f}_h(x))dx,$$

based on the sum of the integrated squared bias and the variance. The bandwidth h is specified as :

$$h = CQn^{-\frac{1}{2}},$$

where C is the kernel-option. If Q is the interquartile range, and n is the sample size, then C is related to h by the previous formula. We considered $C = MISE$.

An approximation is provided by :

$$AMISE(h) = \frac{1}{4}h^4 \left(\int_t t^2 k(t)dt \right)^2 \int_x (f''(x))^2 dx + \frac{1}{nh} \int_t k(t)^2 dt.$$

Plugging in the estimator of equation (7.11) in equation (7.10), we get the conditional FDH efficiency measure for the output oriented case :

$$\hat{\lambda}_n(x, y|z) = sup \left\{ \lambda | \hat{F}_Y(\lambda y|x, z) > 0 \right\}. \tag{7.12}$$

Simar and Daraio remember that the asymptotic properties for this estimator have not yet been derived.

7.3 Empirical Results

In this application only the continuous variable nonperforming loans (q) was analysed. The discrete ones were not considered since to calculate efficiencies based on the probabilistic approach, it would be necessary to divide the data set in so many groups as provided by the combinations of the levels of each variable. At the end, the subsamples are too small. The main routines to compute the probabilistic measures where gently provided by professor Simar and Cinzia Daraio. They were implemented using MATLAT ([1]).

Since the main interest is to investigate the influence of nonperforming loans on bank efficiency, it is calculated the unconditional and conditional probabilistic efficiency measures, $\hat{\lambda}_n(x,y)$ and $\hat{\lambda}_n(x,y|q)$ respectively. Differences on them indicate that nonperforming loans do influence the process. A graph of the rank of their ratio ($\frac{\hat{\lambda}_n(x,y|q)}{\hat{\lambda}_n(x,y)}$) versus the rank of nonperforming loans is available in Figure A.1.

The bandwidth selection method (k-Nearest Neighbor-KNN) suggested by Daraio and Simar [27] required to obtain the nonparametric estimator of $F(x|y,q)$ was not adequate to this data set. The number of observations k provided by this method was either the full sample size or only one observation. The mean integrated square error method (MISE), which value was minimized by the quadratic kernel with bandwidth $h = 0.5308$, is chosen.

To evaluate the relationship between the efficiency score and nonperforming loans, it is calculated the Spearman rank correlation between $\frac{\hat{\lambda}_n(x,y|q)}{\hat{\lambda}_n(x,y)}$ and q (-0.32, p-value=0.0019), significant at the 1% level. Based on the Kolmogorov-Smirnov two-sample test (Table A.27) it is evaluated if the empirical distribution coincides with the expected distribution assuming $\hat{\lambda}_n(x,y) = \hat{\lambda}_n(x,y|q)$. At the 1% level the distributions differ.

Spearman rank correlation and Kolmogorov-Smirnov test indicate that nonperforming loans influence the production process. The negative correlation means that the contextual variable (nonperforming loans) corresponds to an unfavorable factor to the response. The efficiency level decreases as nonperforming loans increases. Also a regression model was applied (Table A.28), confirming previous results. Assuming a gamma distribution, also nonperforming loans is significant, and as it increases, the level of efficiency decreases.

Chapter 8

Conclusions

The thesis extends Banker's results [7] that fundaments a formal statistical basis for the efficiency evaluation techniques of DEA. It is demonstrated the strong consistency of the DEA estimator of a monotone increasing and concave production function, relaxing the assumption of identically distributed inefficiencies. This desirable asymptotic property justifies inference in a two-stage approach that models effects causing inefficiency.

Small samples results are inspected by Monte Carlo simulation. Inefficiencies are estimated based on a univariate production model assuming original inefficiencies uncorrelated. Since the observed correlations were not significant, there is evidence in a production model against Simar and Wilson critics concerning the use of the two-stage approach. They argue that estimated DEA efficiencies are correlated and consequently inference in the two-stage approach may be invalid.

Techniques are considered to evaluate the influence of some contextual variables on the output oriented efficiency measures of commercial banks in Brazil for the year 2001. Investment securities, total loans and demand deposits are the bank outputs used as a multiple output vector and also combined in a single measure. Labor, loanable funds and stock of physical capital are the bank inputs. The significance of the following technical effects is evaluated : bank nature, bank type, bank size, bank control, bank origin and nonperforming loans. Here, specific results for the technical effects that classify the banks will be omitted, since the main interest is on the influence of nonperforming loans on the level of efficiency.

The thesis contributes to the literature suggesting competing statistical models that are justified in terms of the stochastic properties of the production responses in the DEA context. These models are presented in Chapter 4 . The range of model alternatives include the use of nonparametric analysis of covariance, the fit of the truncated normal and the exponential distribution and a general class of Tobit models allowing for heteroscedasticity, fit via maximum likelihood. Conditional moment test of specification is a new alternative suggested that confirm the adequacy of the models.

In Chapter 5 a parametric analysis of covariance is applied and its adequacy is checked by a randomization process with the aim of checking models assumptions.

It is demonstrated, by Monte Carlo simulation, that the restrictions highlighted by Simar and Wilson [42] can not be generalized. Inference on the two-stage approach is formally justified relaxing the assumption of independent identically distributed inefficiencies in production models. Bootstrap procedures are applied with the aim of confirming and comparing asymptotic results. Neither the simple bootstrap algorithm corrected for the bias, nor the algorithms suggested by Simar and Wilson showed different results concerning the significance of nonperforming loans. This variable does not influence the efficiency level.

The previous empirical analysis are based on a two-stage approach where first a nonparametric DEA efficiency measure is obtained and then the 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. In none of these models nonperforming loans appeared to have significant influence on the level of efficiency of brazilian banks.

A different result was obtained for the probabilistic approach, explored in Chapter 7. There is evidence that nonperforming loans do influence other efficiency measures. The conclusion is based on the Spearman rank correlation between the ratio of the conditional probabilistic measure to the unconditional and nonperforming loans. Also the Kolmogorov-Smirnov two-sample test is considered to compare the empirical distribution with the ratio of the conditional to unconditional probabilistic distribution function. There is evidence that the two distributions differ. The observed negative correlation means that the contextual variable (nonperforming loans) corresponds to an unfavorable factor to the response. The same conclusion arrived when one regresses the ratio against nonperforming loans.

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.

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 be studied using our methodology, and we would expect that results would change depending on specific 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 extension of this work is to apply a similar analysis to a panel data and verify not only the variables that influence the production process, but also if changes occurred during this period. Institutions that supervise the banking system have main interest in following bank's performance. The literature suggests the use of Malmquist indices and respective decompositions that usually involve ratios of distance functions, following the lines suggested by Fare and Grosskopf ([30] and [31]).

References

[1] Matlab : The language of technical computing. The MathWorks, Inc, Version 7.0.4.365 (R14) Service Pack 2, License Number : 215808, Banco Central do Brasil.

[2] Sas for windows. SAS 9.1.3 Service Pack 3.

[3] J. N. Boles (1966). Efficiency squared - efficiency computation of efficiency indexes. Proceedings of the 39th Annual Meeting of the Western Farm Economics Association, 137-142.

[4] S. N. Afriat (1972). Efficiency estimation on production functions. *International Economic Review, 13, 568-598.*

[5] B. W. Silverman (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

[6] P. Hall (1986). On the number of bootstrap simulations required to construct a confidence interval. *The Annals of Statistics, 14, 1453-1462.*

[7] R. D. Banker (1993). Maximum likelihood consistency and dea : a statistical foundation. *Management Science, 39(10) :1265-1273.*

[8] G. S. Souza (1998). *Introdução aos Modelos de Regressão Linear e Não-Linear.* Embrapa.

[9] W. J. Conover (1998). *Practical Nonparametric Statistics.* Wiley, NY.

[10] M. Nakane (1999). Productive efficiency in brazilian banking sector. Texto para Discussão 20/99, IPE-USP, São Paulo.

[11] G. S. Souza (2001). Statistical properties of data envelopment analysis estimators of production functions. *Brazilian Journal of Econometrics, 21(2) :291-322.*

[12] M. Sathie (2001). X-efficiency in australian banking : an empirical investigation. *Journal of Banking and Financing 25, 613-630*, 2001.

[13] E. Tortosa-Ausina (2002). Bank cost efficiency and output specification, journal of productive analysis. *Journal of Productive Analysis, 18, 199-222.*

[14] M. B. Campos (2002). Produtividade e eficiência do setor bancário privado brasileiro de 1994 a 1999. Dissertação de Mestrado, EASP-FGV, São Paulo.

[15] P. W. Wilson (2003). Testing independence in models of productive efficiency. *Journal of Productivity Analysis, 20, 361-390.*

[16] W. H. Greene (2003). *Econometric Analysis.* Prentice Hall, 5th ed.

[17] G. S. Souza (2005). Significância de efeitos técnicos na eficiência de produção da pesquisa agropecuária brasileira. *Forthcoming on 'Revista Brasileira de Economia', FGV, Rio.*

[18] R. D. Banker and Natarajan (2001). Evaluating contextual variables affecting productivity using data envelopment analysis. Presented in the Sixth European Workshop on Efficiency and Productivity Analysis.

[19] R. D. Banker and Natarajan (2004). Statistical tests based on dea efficiency scores. Cooper, WW ; Seiford, L.M., Zhu, J. (eds.) Handbook on Data Envelopment Analysis, Kluwer International Series, New York.

[20] R. D. Banker, A. Charnes, and W. W. Cooper (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science, 30, 1078-1092.*

[21] A. N. Berger and R.D. Young (1997). Problem loans and cost efficiency in commercial banks. *Journal of Banking and Finance 21, 849-870.*

[22] A. N. Berger and D. B. Humphrey (2000). *Efficiency of financial institutions : international survey and directions for future research, in Performance of Financial Institutions : Efficiency, Innovation, Regulation.* Cambridge, UK.

[23] C. Cazals, J. P. Florens, and L. Simar (2002). Nonparametric frontier estimation : a robust approaoch. *Journal of Econometrics, 106, 1-25.*

[24] A. Charnes, W. W. Cooper, and E. Rhodes (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research, 2, 429-444.*

[25] T. Coelli, D. S. Rao, and G. E. Battese (1998). An introduction to efficiency and productivity analysis. *Kluwer, Boston.*

[26] R. J. Colwell and E. P. Davis (1992). Output and productivity in banking. *Scandinavian Journal of Economics,94, Supplement, 111-129.*

[27] C. Daraio and L. Simar (2005). Introducing environmental variables in nonparametric frontier models : a probabilistic approach. *Journal of Productivity Analysis, 24, 93-121.*

[28] D. Deprins, L. Simar, and H. Tulkens (1984). Measuring labor efficiency in post offices. in the performance of public enterprises : Concepts and measurements. M. Marchand, P. Pestieau and H. Tulkens (eds.), Amsterdam, North-Holland, 243-267.

[29] R. A. Eisenbeis, G. D. Ferrier, and S. H. Kwan (1999). The informativeness of stochastic frontier and programming frontier efficiency scores : Cost efficiency and other measures of bank holding company performance. Working paper 99-23, Federal Reserve Bank of Atlanta.

[30] R. Fare and S. Grosskopf (1996). Intertemporal production frontiers : With dynamic dea. *Boston : Kluwer Academic Publishers*, 1996.

[31] R. Fare and S. Grosskopf (1998). *Malmquist productivity indexes : A survey of theory and practice.* in R. Fĺare, S. Grosskopf and R. Russell (eds.), Essays in Honor of Sten Malmquist, Dordrecht : Kluwer Academic Publishers.

[32] A. Kneip, B. U. Park, and L. Simar (1998). A note on the convergence of nonparametric dea estimators for production efficiency scores. *Econometric Theory, 14, 783-793.*

[33] A. Korostelev, L. Simar, and A.B. Tsybakov (1995a). Efficient estimation of monotone boundaries. *The Annals of Statistics 23, 476-489.*

[34] A. Korostelev, L. Simar, and A.B. Tsybakov (1995b). On estimation of monotone and convex boundaries. Pub. Inst. Stat. Univ. Paris, XXXIX, 1, 3-18.

[35] S. C. Kumbhakar and A. K. Lovell (2000). *Stochastic Frontier Analysis.* Cambridge University Press.

[36] J. E. Leightner and C. A. K. Lovell (1998). The impact of finance liberalization on the performance of thai banks. *Journal of Economics and Business 50, 115-131.*

[37] T. A. McCarthy and S. Yaisawarng (1993). Technical efficiency in new jersey school districts, in the measurement of productive efficiency. *Oxford University Press, New York.*

[38] B. U. Park, L. Simar, and Ch. Weiner (2000). The fdh estimator for productivity efficiency scores : Asymptotic properties. *Econometric Theory, 16, 855-877.*

[39] T. L. Silva and M. J. Neto (2002). Economia de escala e eficiência nos bancos brasileiros após o real. Estudos Econômicos, 32, 577-620.

[40] L. Simar and P. Wilson (2000). Statistical inference in nonparametric frontier models : The state of the art. *Journal of Productivity Analysis, 13, 49-78.*

[41] L. Simar and P. Wilson (2005). Statistical inference in nonparametric frontier models : recent developments and perspectives. forthcoming in The Measurement of Productive Efficiency, 2nd Edition, Harold Fried, C.A.Knox Lovell and Shelton Schmidt, editors, Oxford University Press.

[42] L. Simar and P. Wilson (2007). Estimation and inference in two-stage, semiparametric models of production process. Forthcoming in the Journal of Econometrics 136, 31–64.

[43] G. S. Souza and R. B. Staub (2006). Two stage inference using dea efficiency measurements in univariate production models. *Forthcoming on 'International Transactions of Operations Research'*.

[44] G. S. Souza, B. Tabak, and R. B. Staub (2006). Assessing the significance of factors effects in output oriented dea measures of efficiency : An application to brazilian banks. *Forthcoming on 'Revista Brasileira de Economia de Empresas'*.

[45] D. C. Wheelock and P. W. Wilson (2003). Robust nonparametric estimation of efficiency and technical change in u.s. commercial banking. Federal Reserve Bank of St. Louis, Working Paper 2003-037A.

[46] T. H. Wonacott and R. J. Wonacott (1990). *Introductory Statistics for Business and Economics*. 4th ed, Wiley, New York.

Appendix A

Tables

Variable	Level	N	Mean	L	U
Bank Nature	Commercial	12	0.462	0.267	0.657
	Multiple	81	0.378	0.317	0.44
Bank Type	Credit	33	0.408	0.32	0.496
	Business	24	0.526	0.405	0.646
	Bursary	3	0.746	0	1
	Retail	34	0.508	0.487	0.056
Bank Size	Large	18	0.317	0.181	0.452
	Medium	30	0.386	0.27	0.502
	Small	25	0.419	0.316	0.522
	Micro	21	0.409	0.28	0.538
Bank Control	Private	79	0.405	0.341	0.469
	Public	15	0.288	0.16	0.415
Bank Origin	Foreign	28	0.404	0.291	0.516
	Domestic	66	0.379	0.311	0.447

TAB. A.1 – Descriptive statistics for categorical variables. Response is $1/\phi_j^*$ for a model with combined output y_c. L and U are lower and upper 95% confidence limits.

Variable	Level	N	Mean	L	U
Bank Nature	Commercial	12	0.633	0.466	0.8
	Multiple	81	0.585	0.525	0.646
Bank Type	Credit	33	0.642	0.559	0.726
	Business	24	0.646	0.531	0.761
	Bursary	3	0.75	0	1
	Retail	34	0.478	0.384	0.572
Bank Size	Large	18	0.522	0.388	0.655
	Medium	30	0.528	0.422	0.633
	Small	25	0.634	0.527	0.741
	Micro	21	0.674	0.553	0.795
Bank Control	Private	79	0.594	0.533	0.654
	Public	15	0.555	0.39	0.72
Bank Origin	Foreign	28	0.534	0.419	0.65
	Domestic	66	0.61	0.546	0.674

TAB. A.2 – Descriptive statistics for categorical variables. Response is $1/\phi_j^*$ for a model with multiple output. L and U are lower and upper 95% confidence limits.

Variable	ϵ^*	ϕ^{*1}	ϕ^{*2}
ϵ^*	1	0.412	0.527
ϕ^{*1}	-	1	0.798
ϕ^{*2}	-	-	1

TAB. A.3 – Rank correlation between DEA residuals ϵ^*, combined output DEA ϕ^{*1} and multiple output DEA ϕ^{*2}.

Variable	Runs	z	p-value
ϵ^*	43	-1.037	0.230
ϕ^{*1}	43	-1.037	0.230
ϕ^{*2}	44	-0.830	0.407

TAB. A.4 – Runs test for DEA residuals ϵ^*, combined output DEA ϕ^{*1} and multiple output DEA ϕ^{*2}.

Source	df	Sum of Squares	Mean Square	F	p-value
Model	11	16,676.60	1,516.05	2.37	0.014
Bank Nature	2	1,742.18	871.09	1.36	0.262
Bank Type	3	1,612.79	537.598	0.84	0.476
Bank Size	3	2,910.93	970.31	1.52	0.217
Bank Control	1	1,565.25	1,565.25	2.45	0.122
Bank Origin	1	2,175.47	2,175.47	3.4	0.069
q	1	95.029	95.029	0.15	0.701
Error	82	52,488.90	640.109	-	-
Total	93	69,165.50	-	-	-

TAB. A.5 – Nonparametric analysis of covariance for DEA residuals.

Source	df	Sum of Squares	Mean Square	F	p-value
Model	11	26,438.22	2,403.48	4.61	$<$0.001
Bank Nature	2	2,282.39	1,141.20	2.19	0.118
Bank Type	3	17,049.21	5,683.07	10.91	$<$0.001
Bank Size	3	2,167.12	722.374	1.39	0.253
Bank Control	1	198.099	198.099	0.38	0.539
Bank Origin	1	67.824	67.824	0.13	0.719
q	1	565.008	565.008	1.08	0.301
Error	82	42,723.78	521.022	-	-
Total	93	69,162.00	-	-	-

TAB. A.6 – Nonparametric analysis of covariance for DEA measurements computed for a combined output.

Source	df	Sum of Squares	Mean Square	F	p-value
Model	11	12,083.20	1,098.47	1.59	0.118
Bank Nature	2	1,856.74	928.372	1.34	0.267
Bank Type	3	4,829.72	1,609.91	2.33	0.081
Bank Size	3	1,292.42	430.807	0.62	0.602
Bank Control	1	325.078	325.078	0.47	0.495
Bank Origin	1	1,445.03	1,445.03	2.09	0.152
q	1	104.035	104.035	0.15	0.699
Error	82	56,715.31	691.65	-	-
Total	93	68,798.50	-	-	-

TAB. A.7 – Nonparametric analysis of covariance for DEA measurements computed for a multiple output.

Model	-2ll	Parms	AIC	BIC
Truncated Normal	167.7	13	193.7	226.8
Exponential	196.2	12	220.2	250.7
Tobit (at zero)	210.1	13	236.1	269.1
Heteroscedastic Tobit (at zero)	206.2	16	238.3	279.0

TAB. A.8 – Parametric models for DEA residuals ϵ^*. -2ll is twice the log-likelihood, Parms is the number of parameters and AIC and BIC are the Akaike and Schwarz information criteria, respectively.

Model	-2ll	Parms	AIC	BIC
Tobit (at 1)	429.3	13	455.3	488.4
Heteroscedastic Tobit (at 1)	413.6	16	445.6	486.3
Truncated Normal (Tobit at 1)	408.5	13	434.5	467.6
Gamma (Tobit at 1)	393.4	13	419.4	452.5
Exponential (Tobit at 1)	436.0	12	460	490.5

TAB. A.9 – Parametric models for DEA responses for combined output ϕ^{*1}. -2ll is twice the log-likelihood, Parms is the number of parameters and AIC and BIC are the Akaike and Schwarz information criteria, respectively.

Model	-2ll	Parms	AIC	BIC
Tobit (at 1)	328.7	13	354.7	387.8
Heteroscedastic Tobit (at 1)	307.2	16	339.2	379.2
Truncated Normal (Tobit at 1)	310.8	13	336.8	369.9
Gamma (Tobit at 1)	296.3	13	317.3	350.4
Exponential (Tobit at 1)	428.3	12	452.3	482.8

TAB. A.10 – Parametric models for DEA responses for multiple output ϕ^{*2}. -2ll is twice the log-likelihood, Parms is the number of parameters and AIC and BIC are the Akaike and Schwarz information criteria, respectively.

Variable	Estimate	Standard Error	t	p-value
Intercept	0.012	1.412	0.01	0.993
n_1	-1.299	1.412	-0.92	0.360
n_2	-0.491	1.327	-0.37	0.712
t_1	0.411	0.483	0.85	0.397
t_2	0.258	0.383	0.67	0.502
t_3	-0.827	0.908	-0.91	0.365
s_1	1.136	0.606	1.87	0.064
s_2	0.846	0.481	1.76	0.082
s_3	0.895	0.437	2.05	0.043
c_1	0.914	0.535	1.71	0.091
o_1	-0.567	0.278	-2.04	0.044
q	0.040	0.076	0.52	0.605
σ^2	0.788	0.198	3.97	$<$0.001

TAB. A.11 – Parametric model for DEA residuals ϵ^*. Truncated normal distribution.

Variable	Estimate	Standard Error	t	p-value
Intercept	1.773	0.673	2.63	0.010
n_1	-0.823	0.624	-1.32	0.190
n_2	-0.617	0.606	-1.02	0.311
t_1	-1.067	0.240	-4.44	<0.001
t_2	-1.144	0.198	-5.79	<0.001
t_3	-1.752	0.397	-4.41	<0.001
s_1	-0.778	0.276	-2.82	0.006
s_2	-0.284	0.217	-1.31	0.193
s_3	-0.051	0.198	-0.26	0.797
c_1	0.238	0.200	1.19	0.237
o_1	-0.167	0.150	-1.11	0.269
q	-0.046	0.039	-1.18	0.243
P	3.079	0.463	6.65	<0.001

TAB. A.12 – Parametric model for DEA measurements from combined output ϕ^{*1}. Tobit with censoring at 1, gamma distribution with shape parameter P.

Variable	Estimate	Standard Error	t	p-value
Intercept	1.135	0.687	1.65	0.012
n_1	-0.967	0.635	-1.50	0.132
n_2	-0.860	0.615	-1.40	0.165
t_1	-0.588	0.245	-2.40	0.018
t_2	-0.523	0.204	-2.56	0.012
t_3	-0.938	0.411	-2.28	0.025
s_1	-0.354	0.279	-1.27	0.208
s_2	-0.015	0.22	-0.07	0.944
s_3	0.127	0.199	0.64	0.525
c_1	0.119	0.210	0.57	0.572
o_1	-0.395	0.153	-2.58	0.011
q	-0.024	0.042	-0.58	0.565
P	2.976	0.486	6.12	<0.001

TAB. A.13 – Parametric model for DEA measurements from multiple output ϕ^{*2}. Tobit with censoring at 1, gamma distribution with shape parameter P.

Response Model	-2ll	ϵ^* LR	p-value	-2ll	ϕ^{*1} LR	p-value	-2ll	ϕ^{*1} LR	p-value
Full	167.710			291.338			393.406		
Bank Nature	170.662	2.953	0.228	294.135	2.798	0.247	395.882	2.476	0.290
Bank Type	170.550	2.840	0.417	299.671	8.333	0.040	424.586	31.180	$<$0.001
Bank Size	173.551	5.441	0.142	295.483	4.145	0.246	402.648	9.242	0.026
Bank Control	171.099	3.390	0.066	291.656	0.318	0.573	394.783	1.376	0.241
Bank Origin	171.950	4.240	0.039	297.887	6.549	0.010	394.655	1.249	0.264
q	167.973	0.264	0.607	291.655	0.317	0.573	394.649	1.243	0.265

TAB. A.14 – Likelihood ratio test statistic -LR for the effects of interest. -2ll is twice the log-likelihood. ϵ^* is the DEA residual. ϕ^{*1} and ϕ^{*2} are DEA measurements for combined and multiple outputs respectively.

Sample	Percentile (%)	Quantile F(n,n)	Empirical Percentile	
			$\mu = 300$	$\mu = 600$
30	99	2.39	98.87	98.47
	95	1.84	93.13	93.33
	90	1.61	88.40	87.93
90	99	1.64	98.73	98.80
	95	1.42	94.27	94.13
	90	1.31	89.07	88.47
150	99	1.46	98.87	98.87
	95	1.31	94.80	94.80
	90	1.23	89.80	89.80

TAB. A.15 – Empirical percentiles for group comparisons when residuals are generated independently from exponential distributions with means $\mu = 300$ and $\mu = 600$.

Sample	Percentile (%)	Quantile F(n/2,n/2)	Empirical Percentile	
			$\mu = 300$	$\mu = 600$
30	99	3.52	97.60	97.73
	95	2.40	92.80	92.93
	90	1.97	86.80	87.33
90	99	2.02	98.73	98.73
	95	1.64	94.33	94.33
	90	1.47	89.73	89.73
150	99	1.72	99.20	99.20
	95	1.47	94.47	94.47
	90	1.35	88.93	88.93

TAB. A.16 – Empirical percentiles for group comparisons when residuals are generated independently from half normal distributions with means $\mu = 300$ and $\mu = 600$.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	2.0830	0.1894	2.98	0.0022
Bank Nature	2	0.1643	0.0822	1.29	0.2797
Bank Type	3	1.6268	0.5423	8.54	<.0001
Bank Size	3	0.2603	0.0868	1.37	0.2588
Bank Control	1	0.0042	0.0042	0.07	0.7974
Bank Origin	1	0.0313	0.0313	0.49	0.4843
q	1	0.0657	0.0657	1.04	0.312
Error	82	5.2065	0.0635	-	-
Total	93	7.2894	-	-	-

TAB. A.17 – Parametric analysis of covariance for DEA measurements computed for a combined output

Variable	p-value (model)	p-value (simulation)
Bank Nature	0.2797	0.2865
Bank Type	<.0001	0.0001
Bank Size	0.2588	0.2624
Bank Control	0.7974	0.7946
Bank Origin	0.4843	0.4894
q	0.3120	0.4425

TAB. A.18 – P-values of the parametric analysis of covariance on a combined output and respective p-values of the simulation for each variable

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	1.0404	0.0946	1.32	0.2305
Bank Nature	2	0.1372	0.0686	0.95	0.3891
Bank Type	3	0.4734	0.1578	2.2	0.0948
Bank Size	3	0.1170	0.0390	0.54	0.6546
Bank Control	1	0.0327	0.0327	0.45	0.502
Bank Origin	1	0.1180	0.1180	1.64	0.2036
q	1	0.0134	0.0134	0.19	0.6672
Error	82	5.8932	0.0719	-	-
Total	93	6.9336	-	-	-

TAB. A.19 – Parametric analysis of covariance for DEA measurements computed for a multiple output

Variable	p-value (model)	p-value (simulation)
Bank Nature	0.3891	0.4336
Bank Type	0.0948	0.0955
Bank Size	0.6546	0.6479
Bank Control	0.5020	0.5065
Bank Origin	0.2036	0.2010
q	0.6672	0.7538

TAB. A.20 – P-values of the parametric analysis of covariance on a multiple output and respective p-values of the simulation for each variable

Variable	Mean	Std	Skew	Kurt	KS	p-value	Rel Bias	Z	P(Z>z)
Intercept	-1.15	4.67	-4.80	28.55	0.25	0.01	-9411.55	31.80	0.00
n_1	0.01	4.63	4.74	27.78	0.25	0.01	-101.04	36.00	0.00
n_2	0.79	4.61	4.87	29.19	0.25	0.01	-260.76	37.40	0.00
t_1	0.38	0.47	-0.01	0.22	0.01	0.15	-6.94	2.29	0.01
t_2	0.24	0.37	0.00	0.22	0.01	0.15	-8.12	2.12	0.02
t_3	-1.02	1.20	-2.60	16.04	0.11	0.01	23.63	8.33	0.00
s_1	1.06	0.57	0.15	0.08	0.02	0.15	-6.24	4.53	0.00
s_2	0.79	0.44	0.13	0.15	0.01	0.15	-6.59	4.49	0.00
s_3	0.84	0.39	0.29	0.41	0.03	0.01	-5.71	4.54	0.00
c_1	0.88	0.52	0.44	0.54	0.04	0.01	-3.57	2.36	0.01
o_1	-0.54	0.27	-0.07	0.01	0.02	0.07	-4.03	3.19	0.00
q	0.03	0.08	-0.27	0.59	0.03	0.01	-17.63	3.55	0.00
σ^2	0.66	0.18	1.01	1.90	0.08	0.01	-16.03	24.68	0.00

TAB. A.21 – Bootstrap mean, standard error, skewness and kurtosis. Kolmogorov-Smirnov (KS) test for normality, relative bias and its significance test.

Param.	Lower Boot. CI	Upper Boot. CI	Boot Mean	Param. Model	Lower Model CI	Upper Model CI	Bias	Bias Corrected
Intercept	-11.02	2.36	-1.15	0.01	-2.79	2.82	-1.16	1.17
n_1	-3.99	4.62	0.01	-1.30	-4.10	1.50	1.31	-2.61
n_2	-2.82	6.84	0.79	-0.49	-3.13	2.14	1.28	-1.77
t_1	-0.47	1.34	0.38	0.41	-0.55	1.37	-0.03	0.44
t_2	-0.44	0.97	0.24	0.26	-0.50	1.02	-0.02	0.28
t_3	-3.69	0.62	-1.02	-0.83	-2.63	0.98	-0.20	-0.63
s_1	0.12	2.39	1.06	1.14	-0.07	2.34	-0.07	1.21
s_2	0.07	1.81	0.79	0.85	-0.11	1.80	-0.06	0.90
s_3	0.23	1.82	0.84	0.90	0.03	1.76	-0.05	0.95
c_1	0.09	2.17	0.88	0.91	-0.15	1.98	-0.03	0.95
o_1	-1.17	-0.07	-0.54	-0.57	-1.12	-0.02	0.02	-0.59
q	-0.12	0.19	0.03	0.04	-0.11	0.19	-0.01	0.05
σ^2	0.58	1.68	0.66	0.79	0.39	1.18	-0.13	0.91

TAB. A.22 – Bootstrap confidence intervals and means, estimated confidence intervals and parameters from the truncated normal model, bias and bias corrected parameters.

Var	Estimate	Standard Error	t	p-value	Lower	Upper
Intercept	1.358	0.562	2.41	0.018	0.24	2.48
n_1	-0.673	0.514	-1.31	0.194	-1.69	0.35
n_2	-0.555	0.497	-1.12	0.267	-1.54	0.43
t_1	-1.183	0.203	-5.84	$<.0001$	-1.59	-0.78
t_2	-1.098	0.170	-6.45	$<.0001$	-1.44	-0.76
t_3	-1.476	0.380	-3.89	0.000	-2.23	-0.72
s_1	-0.857	0.240	-3.58	0.001	-1.33	-0.38
s_2	-0.433	0.186	-2.32	0.023	-0.80	-0.06
s_3	-0.162	0.172	-0.94	0.348	-0.50	0.18
c_1	0.290	0.170	1.7	0.092	-0.05	0.63
o_1	-0.095	0.130	-0.73	0.467	-0.35	0.16
q	-0.019	0.036	-0.54	0.592	-0.09	0.05
P	4.635	0.683	6.79	$<.0001$	3.28	5.99

TAB. A.23 – Parametric model for DEA measurements from combined output ϕ^{*1}, excluding $\hat{\phi}^{*1} = 1$. Tobit with censoring at 1, gamma distribution with shape param. P

Var	Bootstrap	without	$\hat{\phi}^{*1} = 1$	Model		
	Mean	Low	Upper	Mean	Low	Upper
Intercept	1.61	0.40	2.59	1.77	0.42	3.12
n_1	-0.74	-1.62	0.51	-0.82	-2.06	0.41
n_2	-0.63	-1.47	0.51	-0.62	-1.82	0.59
t_1	-1.18	-1.61	-0.77	-1.07	-1.54	-0.59
t_2	-1.09	-1.43	-0.78	-1.14	-1.54	-0.75
t_3	-1.39	-2.25	-0.71	-1.75	-2.54	-0.97
s_1	-0.85	-1.35	-0.34	-0.78	-1.33	-0.23
s_2	-0.43	-0.81	-0.04	-0.28	-0.71	0.15
s_3	-0.16	-0.49	0.16	-0.05	-0.44	0.34
c_1	0.28	-0.06	0.64	0.24	-0.16	0.63
o_1	-0.10	-0.34	0.15	-0.17	-0.46	0.13
q	-0.01	-0.10	0.05	-0.05	-0.12	0.03
P	3.74	3.32	5.40	3.08	2.16	4.00

TAB. A.24 – Algorithm 1 - Bias corrected bootstrap means and percentile confidence intervals with and without estimated efficiencies equal to 1 ($\hat{\phi}^{*1} = 1$) and parameters of the Tobit model with censoring at 1, gamma distribution with shape param. P and respective confidence intervals

Var	Bootstrap	without	$\hat{\phi}^{*1} = 1$	Model		
	Mean	Low	Upper	Mean	Low	Upper
Intercept	1.63	0.25	2.83	1.77	0.42	3.12
n_1	-0.70	-1.85	0.62	-0.82	-2.06	0.41
n_2	-0.47	-1.57	0.83	-0.62	-1.82	0.59
t_1	-1.19	-1.62	-0.73	-1.07	-1.54	-0.59
t_2	-1.23	-1.59	-0.87	-1.14	-1.54	-0.75
t_3	-1.83	-2.55	-1.14	-1.75	-2.54	-0.97
s_1	-0.89	-1.38	-0.39	-0.78	-1.33	-0.23
s_2	-0.39	-0.76	-0.01	-0.28	-0.71	0.15
s_3	-0.13	-0.46	0.22	-0.05	-0.44	0.34
c_1	0.27	-0.12	0.64	0.24	-0.16	0.63
o_1	-0.09	-0.35	0.19	-0.17	-0.46	0.13
q	-0.05	-0.14	0.02	-0.05	-0.12	0.03
P	4.29	3.16	5.78	3.08	2.16	4.00

TAB. A.25 – Algorithm 2 - Double bootstrap means and percentile confidence intervals with and without estimated efficiencies equal to 1 ($\hat{\phi}^{*1} = 1$) and parameters of the Tobit model with censoring at 1, gamma distribution with shape param. P and respective confidence intervals

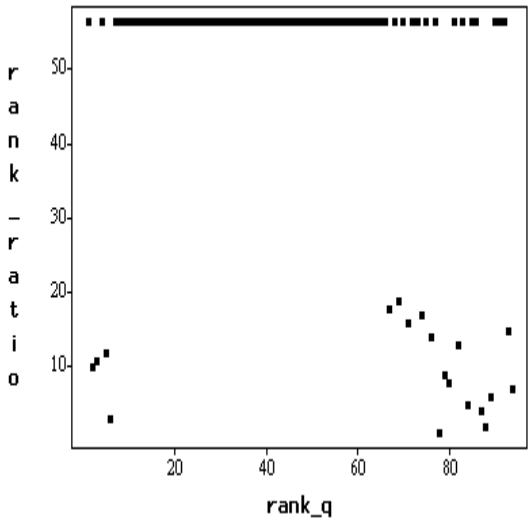

FIG. A.1 – Rank of $\frac{\hat{\lambda}_n(x,y|q)}{\hat{\lambda}_n(x,y)}$ against rank of nonperforming loans (q)

	Correlation	Significance
Spearman	-0.317	0.0019
Pearson correlation	-0.362	0.0003

TAB. A.26 – Pearson and Spearman rank correlation between $\frac{\hat{\lambda}_n(x,y|q)}{\hat{\lambda}_n(x,y)}$ and nonperforming loans and respective levels of significance

Statistic	Value
KS	0.101
D	0.202
KSa	1.386
Pr > KSa	0.043

TAB. A.27 – Asymptotic Kolmogorov-Smirnov two sample test

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

TAB. A.28 – Parametric model for the regression of the ratio $\frac{\hat{\lambda}_n(x,y|q)}{\hat{\lambda}_n(x,y)}$ on q, assuming a gamma distribution with shape parameter P