symmetry

MDPI

Article

Robust Semi-Parametric Inference for Two-Stage Production Models: A Beta Regression Approach

Raydonal Ospina [1,2], Samuel G. F. Baltazar [1], Víctor Leiva [3,*], Jorge Figueroa-Zúñiga [4] and Cecilia Castro [5]

[1] Department of Statistics, CASTLab, Universidade Federal de Pernambuco, Recife 50670-901, Brazil; raydonal@de.ufpe.br (R.O.); samuel.gfbaltazar@gmail.com (S.G.F.B.)
[2] Department of Statistics, IME, Universidade Federal da Bahia, Salvador 40170-110, Brazil
[3] School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso, Valparaíso 2362807, Chile
[4] Department of Statistics, Universidad de Concepción, Concepción 4070386, Chile; jfigueroaz@udec.cl
[5] Centre of Mathematics, University of Minho, 4710-057 Braga, Portugal; cecilia@math.uminho.pt
* Correspondence: victor.leiva@pucv.cl or victorleivasanchez@gmail.com

Abstract: The data envelopment analysis is related to a non-parametric mathematical tool used to assess the relative efficiency of productive units. In different studies on productive efficiency, it is common to employ semi-parametric procedures in two stages to determine whether any exogenous factors of interest affect the performance of productive units. However, some of these procedures, particularly those based on conventional statistical inference, generate inconsistent estimates when dealing with incoherent data-generating processes. This inconsistency arises due to the efficiency scores being limited to the unit interval, and the estimated scores often exhibit serial correlation and have limited observations. To address such inconsistency, several strategies have been suggested, with the most well-known being an algorithm based on a parametric bootstrap procedure using the truncated normal distribution and its regression model. In this work, we present a modification of this algorithm that utilizes the beta distribution and its regression structure. The beta model allows for better accommodation of asymmetry in the data distribution. Our proposed algorithm introduces inferential characteristics that are superior to the original algorithm, resulting in a more statistically coherent data-generating process and improving the consistency property. We have conducted computational experiments that demonstrate the improved results achieved by our proposal.

Keywords: asymmetry; bootstrapping; data envelopment analysis; decision-making units; efficiency; optimization methods; R software; Simar and Wilson algorithm; statistical consistency

check for updates

Citation: Ospina, R.; Baltazar, S.G.F.; Leiva, V.; Figueroa-Zúñiga, J.; Castro, C. Robust Semi-Parametric Inference for Two-Stage Production Models: A Beta Regression Approach. *Symmetry* **2023**, *15*, 1362. https://doi.org/10.3390/sym15071362

Academic Editors: Chong Wang and Victor A. Eremeyev

Received: 5 June 2023
Revised: 23 June 2023
Accepted: 29 June 2023
Published: 4 July 2023

1. Introduction

The data envelopment analysis (DEA) was introduced in [1] and is a non-parametric mathematical tool employed to assess the relative efficiency of productive units, commonly referred to as decision-making units (DMUs). The DEA [2] relies on linear programming to determine an ideal set of weights from entries (inputs) and products (outputs) that the DMUs utilize in their production process. This allows us to build an observed production frontier composed of the most efficient DMUs. This frontier is a benchmark for other DMUs when establishing scores, admitting constant returns [1] or variables for the scale inefficiencies in DEA [3]. The calculation of production efficiency scores or measures, and the construction of the efficiency frontier, may be considered as an approximation of a true unobserved production frontier [4].

The DEA and its derived techniques have gained significant attention in the scientific literature on efficiency analysis. The DEA has applications in energy (efficiency of power stations, electric warfare plants), execution of public policies (efficiency of schools, universities, hospitals), and industry (efficiency of plants and companies), among others; for more examples of applications, see [5].

On the one hand, as the DEA is a non-parametric method, its main advantage is that it does not require knowledge of the functional form that relates inputs and outputs, simplifying the estimates of the efficiency scores for each DMU. On the other hand, a disadvantage of the DEA is that, when applying statistical inference, its results are not formally correct due to its deterministic nature. Despite this, several studies on efficiency use a two-stage approach, where the relative efficiency is estimated in the first stage. Then, the estimated efficiencies are regressed on covariates to infer which exogenous factors are determinants of inefficiency. As mentioned, such an approach is problematic since, due to the non-parametric nature of the DEA, there is no certainty regarding the data-generating process of this regression. Furthermore, since the DEA efficiency scores are limited to the unit interval, employing traditional regression estimators, such as ordinary least squares, is inappropriate.

The two-stage DEA technique has been widely used in the scientific literature of the area [6–10]. In [4], an algorithm and other two-stage DEA procedures were proposed based on an efficiency measure contained in the interval from one to infinity. The algorithm proposed in this work is based on an efficiency score limited to the unit interval. Additionally, in [4], an extensive list of studies that apply two-stage DEA techniques was mentioned. However, conventional statistical inference is often problematic due to its non-parametric nature and the fact that scores almost always show serial correlation. The first attempts to improve the two-stage DEA estimation was proposed in [11], who used the logarithm transformation on the DEA scores to formulate a regression that was followed by some authors utilizing other transformations, as logit or lognormal. Another adaptation that became standard in the area is to use the Tobit model [12] in the second stage, based on the truncated distribution at one, since it avoids problems with predictions outside the original data range.

In [13], the authors compared several regression models employed in the second stage of the DEA and showed that the Tobit structure performs better than the traditional linear regression (based on ordinary least squares) and the inflated beta regression [14]. Although the Tobit model is often sufficient for describing the estimated efficiency scores against exogenous variables [13], a problem remains. The DEA efficiency scores exhibit serial correlation [4], aside from the fact that problems of this type often have small samples, which makes it difficult to utilize traditional inference based on asymptotic approximations. Computationally intensive methods to deal with two-stage DEA have appeared, almost always based on bootstrapping [15]. Some works [16] proposed a bootstrap algorithm that resamples DMUs, dealing with the empirical distribution of data (naive bootstrap). Nonetheless, this algorithm has been proven to be inconsistent [17].

Additionally, in [4], two parametric bootstrap algorithms were proposed (from now, we refer to them as the Simar–Wilson –SW– algorithm) based on single and double bootstrapping and the Tobit model. The SW algorithm employs the truncated normal distribution at a value equal to one to the left and a regression model. In the SW algorithm, a parametric resampling of the efficiency scores is performed using the vector of parameters for the original exogenous covariates estimated with the original sample. In the same work, the authors demonstrated the consistency property of the estimators generated by the SW algorithm. Nevertheless, they acknowledged that no specific reason suggests the truncated normal distribution to be superior to other distributions in the data-generating process.

The SW algorithm is based on removing the efficient observations (that is, with the efficiency measure equal to one), estimating the truncated normal regression coefficients in the remaining observations, and performing a single or double parameter bootstrapping from the estimates in the original regression. This is obtained by generating random pseudo-samples of the truncated normal distribution at one and re-estimating the regression in the pseudo-samples. Then, sets of bootstrap estimates for each coefficient are reached, allowing the estimation of the parameters and standard errors (SE) of the second-stage model, enabling the construction of confidence intervals and hypothesis tests for the model parameters or estimated efficiency measures.

Although the SW algorithm has been widely studied in this area, it is worth remembering that it is based on a truncated normal distribution. Thus, the hypothesis that the efficiency scores follow this distribution is not always validated in real problems. This is because, in many applications, it is common for the efficiency scores to be concentrated close to an upper limit equal to one, mainly when the number of inputs and outputs is close to the number of DMUs [18]. Hence, the need to employ a more flexible probability distribution naturally arises. Until now, the beta distribution and its associated regression model [19–21] have not been considered in DEA under the perspective of the SW algorithm.

Therefore, the objective of the present investigation is to propose and derive a bootstrap approach inspired by the SW algorithm, but based on the beta distribution and its regression supported at the unit interval. Such an interval of the beta distribution contains the range of efficiency scores. In addition, this distribution is quite flexible, permitting us to better accommodate asymmetry in the data distribution, an aspect commonly presented in problems of this type. The utilization of parametric bootstrapping [22] enables us to minimize the serial correlation in the estimated regressions and increase the inference quality, since the hypothesis tests used are frequently asymptotic and DEA problems often have small samples.

In this study, we have opted to utilize simulated data rather than real data. This decision is justified by several advantages that simulated data offer. Obtaining precise data in real-life situations is often challenging, which can limit the applicability of conventional DEA models. Real data can frequently be bounded (interval), ordinal, and ratio-bounded, thereby constraining the range of possible analysis [23]. The simulated production units can be viewed as an abstract and generic representation, ideal for depicting a hypothetical scenario wherein the efficiency of production units is evaluated based on inputs and outputs. We chose this simulation approach to be able to control and handle the underlying relationships between variables, which, in turn, enables a more profound understanding of their effects on efficiency measures. By employing simulated data, we can conduct controlled experiments, offering valuable insights into the factors that influence the efficiency of production units. This approach aligns with the primary objective of our study, which is to develop and validate a new methodology for efficiency analysis, without being constrained by real data that may be limited in scope or availability.

The remaining sections of this article are organized as follows. In Section 2, we provide background for proposing the new methodology. Section 3 states the new methodology using two algorithms. In Section 4, the performance of the proposed methodology utilizing Monte Carlo simulation methods is evaluated. Concluding remarks and ideas for future research are discussed in Section 5.

2. Background

In this section, we present some preliminary aspects as background to formulate the new methodology proposed in this work.

2.1. Beta Models

In many statistical applications, it is common that variables are limited to the unit interval, that is, the interval $(0,1)$. Standard regression models, such as normal linear regression, are inadequate for modeling this type of variables, since they allow predictions outside the original range. Different strategies have been proposed to model data limited to $(0,1)$, with regression models based on the beta distribution for the response variable being frequently used [14,19,24,25]. Some extensions of these models have also been recently proposed in [26,27]. The beta regression presented in [19] is a model for continuous variables that assumes values in the interval $(0,1)$; see also [28]. In this model, the regression parameters are interpreted in terms of the mean of the dependent variable, utilizing the parametrization of the probability density function (PDF) of the beta distribution in function of the mean and a precision parameter.

Note that the beta PDF is often represented by:

$$f(y;p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, \quad 0 < y < 1,$$

where $p > 0$, $q > 0$, and Γ is the gamma function. In [19], the authors proposed an alternative parametrization, making $\mu = p/(p+q)$ and $\phi = p+q$. Therefore, we have that the beta PDF is now given by:

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad 0 < y < 1,$$

with $0 < \mu < 1$ and $\phi > 0$. Then, we use the notation $Y \sim \text{Beta}(\mu,\phi)$, where μ is the mean and $\mu(1-\mu)/(1+\phi)$ is the variance. The parameter ϕ is known as the precision parameter, because for fixed μ, as ϕ increases, the variance decreases.

Let $Y_1, \dots, Y_n$ be a sample such that $Y_i \sim \text{Beta}(\mu_i, \phi)$, for $i \in \{1, \dots, n\}$. Thus, the beta regression model is defined as:

$$g(\mu_i) = \sum_{t=1}^{k} \beta_t x_{it} = \eta_i, \quad i \in \{1, \dots, n\}, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)^\top$ is a vector of unknown parameters, with $\boldsymbol{\beta} \in \mathbb{R}^k$; $(x_{i1}, \dots, x_{ik})^\top$ is the set of k known values of the covariates; and g: $(0,1) \to \mathbb{R}$ is a link function, which is strictly monotone, twice differentiable, and employed to describe the relationship (linear or non-linear) between the response variable and the covariates. Thus, $\mu_i = g^{-1}(\eta_i)$, where the estimation of parameters $\boldsymbol{\beta}$ and ϕ is usually done by the maximum likelihood method, whereas the significance of the parameters can be stated through a z-test, whose statistic is the square root of the Wald statistic [29]. Hence, to test the null hypothesis $\mathcal{H}_0$: $\beta_j = \beta_j^{(0)}$ against $\mathcal{H}_1$: $\beta_j \neq \beta_j^{(0)}$, we use:

$$Z = \frac{\widehat{\beta}_j - \beta_j^{(0)}}{\text{SE}(\widehat{\beta}_j)}, \quad j \in \{1, \dots, k\}, \tag{2}$$

where $\text{SE}(\widehat{\beta}_j)$ is the square root of the j-th diagonal element of the inverse of the Fisher information matrix, evaluated in the maximum likelihood estimates. Under $\mathcal{H}_0$, the statistic Z defined in (2) is approximately standard normal distributed. The beta regression is a naturally heteroskedastic model based on a flexible distribution, which can accommodate different types of asymmetries. It is worth noting that this class of models has an approach similar to that of generalized linear models [30]. Observe the different shapes of the beta PDF displayed in Figure 1.

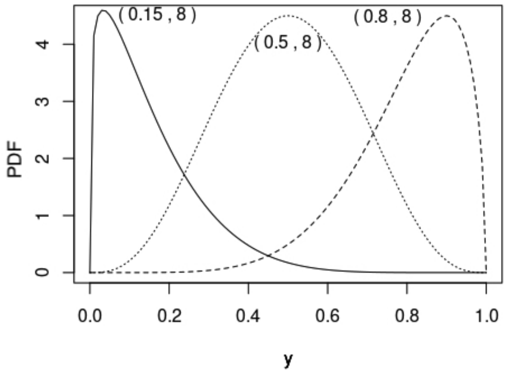

Figure 1. Plots of beta PDFs for the indicated values of μ and $\phi = 8$.

The formulation described in (1) assumes that the precision parameter ϕ is fixed, an assumption that may not always hold true in empirical applications. In [31], an extension of the original model was proposed, in which the mean is described as previously indicated, and the parameter of precision is also similarly formulated through a linear predictor (which may or not employ the same covariates as the mean model) and a link function. Then, we have that:

$$g(\mu_i) = \sum_{j=1}^{k} \beta_j x_{ij} = \eta_i, \quad h(\phi_i) = \sum_{j=1}^{m} \gamma_j z_{ij} = \nu_i, \quad i \in \{1, \ldots, n\},$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m)^\top$ is a vector of unknown parameters in $\mathbb{R}^m$; $(z_{i1}, \ldots, z_{im})^\top$ is the set of m known values of the covariates, which may or may not be equal to the set $(x_{i1}, \ldots, x_{ik})$, with $k + m < n$; and h: $(0, \infty) \to \mathbb{R}$ is a link function to be also strictly monotone, twice differentiable, and employed to describe the relationship (linear or non-linear) between the response variable and the covariates.

The maximum likelihood method estimates the parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and ϕ. However, the functional form of the precision parameter is often unknown, making its modeling difficult. In [32], a robust estimator was proposed for the covariance matrix of the maximum likelihood estimator of $\boldsymbol{\beta}$. This estimator was directly inspired by the sandwich estimators proposed in [33] for the normal linear regression model, later generalized to other regression models [34]. In [32], hypothesis testing on the regression coefficients was performed for the mean submodel, even when ϕ is varying, employing robust estimators of the covariance matrix of $\widehat{\boldsymbol{\beta}}$, eliminating the need of submodels for ϕ. The test statistics, named quasi z, utilized to contrast the null hypothesis $\mathcal{H}_0$: $\beta_j = \beta_j^{(0)}$ against $\mathcal{H}_1$: $\beta_j \neq \beta_j^{(0)}$ are given by:

$$Z_{\mathrm{R}_j} = \frac{\widehat{\beta}_j - \beta_j^{(0)}}{\mathrm{SE}_{\mathrm{R}}(\widehat{\beta}_j)}, \quad j \in \{1, \ldots, k\},$$

where $\mathrm{SE}_{\mathrm{R}}(\widehat{\beta}_j)$ is the square root of the j-th diagonal element of the robust covariance matrix of the estimators, as presented in [32]. Like the z-test, under $\mathcal{H}_0$, Z_{R_j} also has an approximately standard normal distribution.

2.2. Data Envelopment Analysis

To explain the concept of DEA, let us assume the existence of a DMU, denoted by A, and its non-negative sets of inputs $\boldsymbol{x}^A = (x_1^A, \ldots, x_r^A)^\top \in \mathbb{R}_+^r$ and outputs $\boldsymbol{y}^A = (y_1^A, \ldots, y_s^A)^\top \in \mathbb{R}_+^s$, as well as another DMU, denoted by B, with its respective inputs $\boldsymbol{x}^B = (x_1^B, \ldots, x_r^B)^\top \in \mathbb{R}_+^r$ and outputs $\boldsymbol{y}^B = (y_1^B, \ldots, y_s^B)^\top \in \mathbb{R}_+^s$.

Note that the DEA works with the fundamental assumption that if A can produce the quantities of outputs $\boldsymbol{y}_A$, using the quantities of inputs $\boldsymbol{x}_A$, other DMUs could also do the same if they operate efficiently. If DMUs A and B are efficient, they can be linearly combined to create a virtual DMU, which utilizes a mixture of inputs to produce a combination of outputs. The goal of the DEA is to identify the optimal virtual DMU for each DMU in the sample. An inefficient DMU is one where the virtual DMU outperforms the original DMU, either by producing more outputs with the same inputs or by achieving the same outputs with fewer inputs. Similarly, if B produces the quantities of product $\boldsymbol{y}^B$, employing $\boldsymbol{x}^B$ as inputs, then other DMUs might produce the same.

The first model of interest in this work was developed in [1], whose authors constructed the efficiency frontiers assuming constant returns to scale inefficiencies in DEA, known as the DEA-CCR model. Boundaries and, consequently, efficiency scores are calculated using linear programming, in which we look for the set of weights $\boldsymbol{u} = (u_1, \ldots, u_r)^\top$ for the inputs and $\boldsymbol{v} = (v_1, \ldots, v_s)^\top$ for the outputs that maximize the ratio between the linear combination of inputs and the combination of outputs.

Let $\boldsymbol{x}^A = (x_{1A}, \ldots, x_{rA})^\top$ and $\boldsymbol{y}^A = (y_{1A}, \ldots, y_{sA})^\top$ be the set of inputs and outputs of a DMU A, and $\mathcal{P}$ be the set of indices of the DMUs. Then, we can measure the relative efficiency of this DMU, with a focus on inputs, using:

$$\max_{\boldsymbol{u},\boldsymbol{v}} \ h_A = \frac{\sum_{j=1}^{r} u_j y_{jA}}{\sum_{i=1}^{s} v_i x_{iA}},$$

subject to:

$$\frac{\sum_{j=1}^{r} u_j y_{jk}}{\sum_{i=1}^{s} v_i x_{ik}} \leq 1, \quad \forall k \in \mathcal{P},$$
$$v_i, u_j > 0, \quad \forall i, j.$$

The above formulation corresponds to a fractional programming problem requiring a high computational cost to be solved. A possible modification to simplify problem-solving is to fix $\sum_{i=1}^{s} v_i x_{ik} = 1$, for all k. Thus, we have:

$$\max_{\boldsymbol{u},\boldsymbol{v}} \ h_A = \sum_{j=1}^{r} u_j y_{jA},$$

subject to:

$$\sum_{i=1}^{s} v_i x_{ik} = 1,$$
$$\sum_{j=1}^{r} u_j y_{jk} - \sum_{i=1}^{s} v_i x_{ik} \leq 0, \quad \forall k \in \mathcal{P},$$
$$v_i, u_j > 0, \quad \forall i, j.$$

We can calculate the efficiency scores more directly using the dual formulation of the previous problem as:

$$\min_{h_A, \lambda} \ h_A,$$

subject to:

$$h_A x_{iA} - \sum_{k=1}^{n} x_{ik} \lambda_k \geq 0, \quad \forall i,$$
$$-y_{jA} + \sum_{k=1}^{n} y_{jk} \lambda_k \geq 0, \quad \forall j,$$
$$\lambda_k > 0, \quad \forall k \in \mathcal{P}.$$

For variable returns to scale inefficiencies in DEA, in [3], a modification of the above problem was proposed using the formulation given by:

$$\min_{h_A, \lambda} \ h_A,$$

subject to:

$$h_A x_{iA} - \sum_{k=1}^{s} x_{ik} \lambda_k \geq 0, \quad \forall i,$$
$$-y_{jA} + \sum_{k=1}^{r} y_{jk} \lambda_k \geq 0, \quad \forall j,$$
$$\sum_{k=1}^{n} \lambda_k = 1,$$
$$\lambda_k > 0, \quad \forall k \in \mathcal{P}.$$

The addition of the last constraint $\sum_{k=1}^{n} \lambda_k = 1$ introduces convexity at the production frontier. It allows DMUs to operate with either decreasing, increasing, or constant returns to scale inefficiencies in DEA, known as the DEA-BCC model. The estimation of the efficiency measure for the DMU A, which is denoted by $\widehat{\delta}(\boldsymbol{x}_A, \boldsymbol{y}_A)$ or $\widehat{\delta}_A$, is then calculated as $\widehat{\delta}_A = 1/h_A$, where $h_A \geq 1$ is called the "efficiency measure", with the DMU being efficient if $h_A = 1$ and inefficient if $h_A > 1$. Similarly, $0 < \widehat{\delta}_A \leq 1$ is called the "efficiency score", where the DMU is efficient if $\widehat{\delta}_A = 1$ and inefficient if $0 < \widehat{\delta}_A < 1$.

The efficiency calculation for the output-oriented DEA-CCR model is so given by:

$$\max_{h_A, \lambda} h_A,$$

subject to:

$$-h_A y_{jA} + \sum_{k=1}^{s} y_{jk} \lambda_k \geq 0, \quad \forall j,$$

$$x_{iA} + \sum_{k=1}^{r} x_{ik} \lambda_k \geq 0, \quad \forall i, \lambda_k > 0, \quad \forall k \in \mathcal{P}.$$

Therefore, the output-oriented DEA-BCC model is stated as:

$$\max_{h_A, \lambda} h_A,$$

subject to:

$$-h_A y_{jA} + \sum_{k=1}^{s} y_{jk} \lambda_k \geq 0, \quad \forall j,$$

$$x_{iA} + \sum_{k=1}^{r} x_{ik} \lambda_k \geq 0, \quad \forall i,$$

$$\sum_{k=1}^{n} \lambda_k = 1, \quad \lambda_k > 0, \quad \forall k \in \mathcal{P}.$$

In Figure 2, we see an example of an efficiency frontier construction with one input and one output. This figure shows the efficiency frontier constructed assuming variable returns to scale inefficiencies in DEA. The efficient DMUs are located on the efficiency frontier and have an estimated efficiency score $\widehat{\delta}(x_i, y_i) = 1$, which are the benchmarks, that is, the efficiency references for the other DMUs. Hence, the inefficient DMUs may be compared with the virtual DMUs, formed by the linear combination of the efficient DMUs and located on the efficiency frontier.

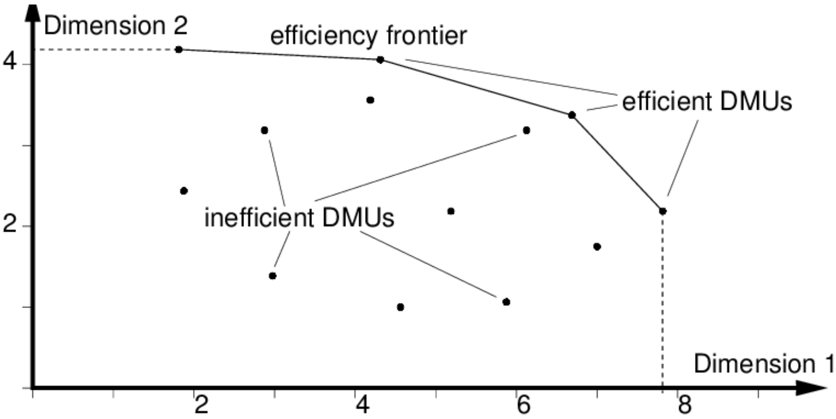

Figure 2. Example of the construction of an empirical efficiency frontier. Source: [35].

3. The New Methodology by Using Two Algorithms

In this section, we propose two simple parametric bootstrap algorithms based on beta regression to estimate a two-stage DEA. These algorithms aim to improve the inference of the exogenous variables of interest in the second stage by employing a distribution that best fits the data. By utilizing bootstrapping, we can enhance the quality and performance of the tests used to detect the significance of the beta regression coefficients, particularly in cases where the DEA efficiency scores exhibit serial correlation and small sample sizes.

3.1. Context

The proposed methodology introduces a new approach for estimating the efficiency measures and assessing the impact of covariates on the efficiency of production units. We conducted Monte Carlo simulations, generating data with known generating processes, to evaluate the performance of the proposed methodology. As mentioned, by using simulated data, we can control and handle the relationships between variables to better understand the effects on efficiency measures. The simulated production units can be considered as an abstract and generic example, representing a hypothetical scenario where the efficiency of production units is evaluated based on inputs and outputs. This approach allows us to conduct controlled experiments and gain insights into the factors that influence the efficiency of production units.

The two bootstrap algorithms presented in this section provide a more robust and reliable approach for estimating the efficiency measures and assessing the impact of covariates. These algorithms, described in detail below, utilize beta regression and employ bootstrapping techniques to improve the quality and performance of the tests used to detect the significance of the beta regression coefficients.

3.2. Bootstrap Algorithm Using Beta Regression with the z-Test

Consider the set of n DMUs with their respective sets $\boldsymbol{x}$ of inputs and $\boldsymbol{y}$ of outputs. In addition, let $\widehat{\delta}_i = \widehat{\delta}(x_i, y_i)$ be the estimated efficiency score for DMU i given the inputs x_i and outputs y_i, for $i \in \{1, \ldots, n\}$.

Let $\boldsymbol{Z}$ be the matrix of covariates, with $\boldsymbol{z}_i$ being its respective rows (set of values of the variables per observation), and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^\top$ be the vector of regression parameters in the second stage. Algorithm 1 describes a simple parametric bootstrap approach based on beta regression to estimate a two-stage DEA.

With the bootstrap p-value and a chosen fixed significance level α, we then have the decision rule in which we reject the null hypothesis of the coefficient of the variable of interest being equal to zero, that is, the variable is not a statistically significant factor for the efficiency if p-value $< \alpha$.

3.3. Bootstrap Algorithm Using Beta Regression with Quasi z-Test

A modification of Algorithm 1 is accomplished by utilizing a quasi z-statistic [32] to avoid possible biases and errors in the inference caused by misspecification in the precision parameter. Algorithm 2 describes a second parametric bootstrap approach based on beta regression to estimate a two-stage DEA considering this modification.

Figure 3 illustrates the underlying flow diagram of Algorithms 1 and 2 as well as their integration into the research methodology. This diagram provides a clear visualization of the information flow, starting from the collection of input and output data from the DMUs, followed by the application of the DEA to estimate efficiency scores. Then, the diagram shows the utilization of the two bootstrap algorithms based on beta regression, along with the evaluation of their performance through computational experiments using Monte Carlo simulations. The diagram also highlights the analysis of the simulation results and the conclusion drawn from them.

Algorithm 1 Approach with the z-test.

```
1: Collect original input and output data from the $n$ DMUs.
2: Use the DEA to estimate the efficiency scores $\widehat{\delta}_i = \widehat{\delta}(x_i, y_i)$, for $i \in \{1, \ldots, n\}$.
3: Employ $m < n$ observations, where $0 < \widehat{\delta}_i < 1$, to estimate $\widehat{\beta}$ such that $\widehat{\delta} = \boldsymbol{Z\beta}$ by the maximum likelihood method utilizing a beta regression, assuming constant precision.
4: Obtain a statistic $c$ of the z-test to contrast significance of the regression coefficients for the covariate of interest.
5: State $\mathcal{H}_0$: $\beta_j = \beta_j^{(0)}$ against $\mathcal{H}_1$: $\beta_j \neq \beta_j^{(0)}$ on the coefficient of covariate $j$ of interest and fit an auxiliary regression $\widehat{\delta} = \boldsymbol{Z\beta}_{\text{aux}}$ also assuming constant precision, that is, $\boldsymbol{\beta}_{\text{aux}} = \boldsymbol{\beta}$, such that $\beta_j = 0$.
6: Enter the bootstrap loop by means of:
     • Generate $m$ random observations $\widehat{\delta}^*$ under the beta distribution, that is, $\widehat{\delta}^* \sim \text{Beta}(\boldsymbol{Z}\widehat{\boldsymbol{\beta}}_{\text{aux}}, \widehat{\phi})$ using the parametrization proposed in [19], where $\boldsymbol{Z}\widehat{\boldsymbol{\beta}}_{\text{aux}}$ are the responses predicted by the regression estimated at Step 5 and $\widehat{\phi}$ is the precision parameter estimated at this same step.
     • Estimate $\widehat{\boldsymbol{\beta}}^*$ such that $\widehat{\delta}^* = \boldsymbol{Z\beta}^*$ with the values of all covariates in matrix $\boldsymbol{Z}$.
     • Calculate the statistic $c^*$ of the z-test for the variable of interest in the previous regression.
     • Repeat the steps $B$ times and count the quantity $q$ of times that $|c^*| \geq |c|$.
7: Calculate the bootstrap $p$-value $p^* = (1 + q)/B + 1$.
8: Make a decision about rejecting $\mathcal{H}_0$ or not at a significance level fixed.
```

Algorithm 2 Approach with the quasi z-test.

```
1: Collect original input and output data from the $n$ DMUs.
2: Use the DEA to estimate the efficiency scores $\widehat{\delta}_i = \widehat{\delta}(x_i, y_i)$, for $i \in \{1, \ldots, n\}$.
3: Employ $m < n$ observations, where $0 < \widehat{\delta}_i < 1$, to estimate $\widehat{\beta}$ such that $\widehat{\delta} = \boldsymbol{Z\beta}$ by the maximum likelihood method utilizing a beta regression, assuming constant precision.
4: Obtain a statistic $d$ of the quasi z-test to contrast significance of the regression coefficients for the covariate of interest.
5: State $\mathcal{H}_0$: $\beta_j = \beta_j^{(0)}$ against $\mathcal{H}_1$: $\beta_j \neq \beta_j^{(0)}$ on the coefficient of covariate $j$ of interest and fit an auxiliary regression $\widehat{\delta} = \boldsymbol{Z\beta}_{\text{aux}}$ also assuming constant precision, that is, $\boldsymbol{\beta}_{\text{aux}} = \boldsymbol{\beta}$, such that $\beta_j = 0$.
6: Enter the bootstrap loop by means of:
     • Generate $m$ random observations $\widehat{\delta}^*$ under the beta distribution, that is, $\widehat{\delta}^* \sim \text{Beta}(\boldsymbol{Z}\widehat{\boldsymbol{\beta}}_{\text{aux}}, \widehat{\phi})$ using the parametrization proposed in [19], where $\boldsymbol{Z}\widehat{\boldsymbol{\beta}}_{\text{aux}}$ are the responses predicted by the regression estimated at Step 5 and $\widehat{\phi}$ is the precision parameter estimated at this same step.
     • Estimate $\widehat{\boldsymbol{\beta}}^*$ such that $\widehat{\delta}^* = \boldsymbol{Z\beta}^*$ with all covariates in matrix $\boldsymbol{Z}$.
     • Calculate the statistic $d^*$ of the quasi z-test for the variable of interest in the previous regression.
     • Repeat the steps of this item $B$ times and count the quantity $q$ of times that $|d^*| \geq |d|$.
7: Calculate the bootstrap $p$-value $p^* = (1 + q)/B + 1$.
8: Make a decision about rejecting $\mathcal{H}_0$ or not at a significance level fixed.
```

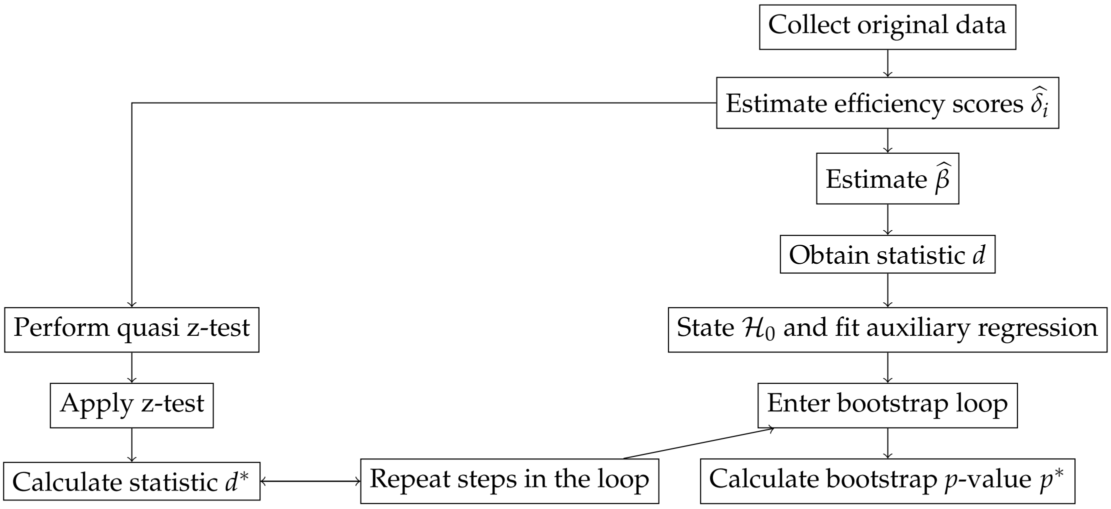
Collect original data
Estimate efficiency scores $\widehat{\delta}_i$
Estimate $\widehat{\beta}$
Obtain statistic d
State $\mathcal{H}_0$ and fit auxiliary regression
Enter bootstrap loop
Calculate bootstrap p-value p^*
Perform quasi z-test
Apply z-test
Calculate statistic d^*
Repeat steps in the loop

Figure 3. Flow diagram of sequential execution with two algorithms for estimation and testing of the proposed DAE methodology.

4. Computational Experiments

In this section, we evaluate the performance of the two inferential approaches proposed in this work. We conduct Monte Carlo simulations similarly to [4]. We generate data with a known generating process and apply the bootstrap algorithms to each of the Monte Carlo replicates. For comparison purposes, we consider the SW double bootstrap algorithm.

4.1. Simulation Setting

In the simulation, we establish the amount n of desired artificial DMUs and carry out the following procedure:

- Simulate data from observations of covariates randomly generated from $Z_{i1} \sim \mathrm{N}(2,4)$, for $i \in \{1,\ldots,n\}$, that is, normally distributed.
- Generate values ε_i from the standard normal distribution truncated at $1 - \beta_0 - z_{1i}\beta_1$ and state $\delta_i = \beta_0 + z_{1i}\beta_1 + \varepsilon_i$, for $i \in \{1,\ldots,n\}$.
- Obtain the set of inputs (considering one input and one output) and obtain values from $X_i \sim \mathrm{U}(6,16)$, for $i \in \{1,\ldots,n\}$, that is, uniformly distributed.
- Set $y_i = \delta^{-1} x_i^{3/4}$ to generate the output data for $i \in \{1,\ldots,n\}$.
- Add a second covariate from $Z_{i2} \sim \mathrm{N}(2,4)$ to the model, which is not part of the data generating process.

In our simulation, we set $\beta_0 = \beta_1 = 0.5$ allowing us to test the significance of the parameters in the linear predictor $\beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2}$. Note that we know in advance that parameters β_0 and β_1 are statistically significant (as they are part of the data-generating process) and β_2 is not (as it is not part of this process).

4.2. Simulation Results

Each algorithm was tested over 5000 Monte Carlo replicates. Within each replicate, Algorithms 1 and 2 were run with 2000 iterations each in the bootstrap loop, as recommended in [36], to minimize power loss without significantly increasing computational cost. The SW algorithm had 2000 repetitions in the first loop and 100 repetitions in the second loop of the bootstrap, as suggested in [4]. We employed the logit link function in Algorithms 1 and 2. We estimated power by evaluating the proportion of correct rejection of the null hypothesis in the test $\mathcal{H}_0$: $\beta_1 = 0$ against $\mathcal{H}_1$: $\beta_1 \neq 0$. We assessed test coverage by evaluating the correct proportion of non-rejection of the null hypothesis $\mathcal{H}_0$: $\beta_2 = 0$ against $\mathcal{H}_1$: $\beta_2 \neq 0$. All tests were conducted at level $1 - \alpha = 0.95$ and the simulations were carried out in a computational environment of the `R` software.

The findings from the results presented in Tables 1 and 2 are consistent across the three quantities of DMUs studied. Under the selected data-generating process, the z-test made with the SW algorithm tends to be conservative, meaning the estimated probability of correctly not rejecting the null hypothesis never exceeds the specified nominal level [37]. In contrast, the significance test in Algorithm 1 leans liberal (Table 2). The quasi z-test conducted with Algorithm 2, which we propose, offers superior test coverage, coming closer to the selected level of $1 - \alpha = 0.95$. Another observation is that, with a smaller number of DMUs ($n = 40$), tests conducted with Algorithms 1 and 2 outperform and yield higher power than the z-tests with the SW algorithm (Table 1), demonstrating a higher rate of correct null hypothesis rejection. As the number of DMUs increases ($n \geq 70$), the tests performed with Algorithm 2 show less power than those with the SW algorithm. The performance of the algorithms across the three quantities of DMUs studied suggests using Algorithm 2 for two-stage DEA involving smaller quantities of DMUs. This is because it displays equal or superior power than the SW algorithm while keeping the test coverage at the specified nominal level.

Table 1. Estimated rejection rate for the null hypothesis $\beta_1 = 0$.

n	Algorithm 1	Algorithm 2	SW Algorithm
40	0.230	0.204	0.191
70	0.433	0.375	0.405
100	0.581	0.520	0.605

Table 2. Estimated non-rejection rate for the null hypothesis $\beta_2 = 0$.

n	Algorithm 1	Algorithm 2	SW Algorithm
40	0.933	0.943	0.967
70	0.932	0.944	0.965
100	0.935	0.950	0.968

5. Conclusions

In this article, we have introduced two parametric bootstrap algorithms that leverage the beta regression model proposed in [19] as an alternative to the Simar and Wilson algorithm for inferring the significance of exogenous variables determining the efficiency of production units. Our simulations show that Algorithm 2 performed better in terms of power and coverage when the number of decision-making units under analysis is smaller ($n \approx 40$), positioning it as a viable alternative for two-stage data envelopment analysis with smaller production unit sets.

Based on the behavior of the algorithms with different quantities of decision-making units, we recommend Algorithm 2 for studying two-stage data envelopment analysis problems with smaller production unit sets. This algorithm exhibits equal or greater power than the Simar and Wilson algorithm, while maintaining the test coverage at the specified nominal level.

For future research, we suggest conducting simulations with more inputs and outputs, as well as fewer decision-making units. It could also be worth considering a quantile regression model based on the Kumaraswamy distribution [38] instead of the beta regression model. This model could be a compelling alternative, especially when outliers are present in the response variable under consideration. Furthermore, exploring the use of inflated beta regression [39] in algorithm formulation, or incorporating heuristics or other optimization algorithms in the maximum likelihood method (steps 3 of both Algorithms 1 and 2) for larger n values, may yield valuable insights. Lastly, we propose developing an R package encompassing the algorithms presented here to facilitate their implementation and use in future studies.

Author Contributions: Data curation, R.O. and S.G.F.B.; formal analysis, R.O., S.G.F.B., V.L., J.F.-Z. and C.C.; investigation, R.O., S.G.F.B., V.L., J.F.-Z. and C.C.; methodology, R.O., S.G.F.B., V.L., J.F.-Z. and C.C.; writing—original draft, R.O., S.G.F.B. and J.F.-Z.; writing—review and editing, V.L. and C.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by the National Council for Scientific and Technological Development (CNPq) through the grant 303192/2022-4 (R.O.); and Comissão de Aperfeiçoamento de Pessoal do Nível Superior (CAPES), from the Brazilian government; and by FONDECYT, grant number 1200525 (V.L.), from the National Agency for Research and Development (ANID) of the Chilean government under the Ministry of Science and Technology, Knowledge, and Innovation.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data and codes used in this work are available under request.

Acknowledgments: The authors would also like to thank the Editor and three Reviewers for their constructive comments, which led to the improvement of this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Charnes, A.; Cooper, W.W.; Rhodes, E. Measuring the efficiency of decision making units. *Eur. J. Oper. Res.* **1978**, *2*, 429–444. [CrossRef]
2. Wong, W.P. A global search method for inputs and outputs in data envelopment analysis: Procedures and managerial perspectives. *Symmetry* **2021**, *13*, 1155. [CrossRef]
3. Banker, R.D.; Charnes, A.; Cooper, W.W. Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Manag. Sci.* **1984**, *30*, 1078–1092. [CrossRef]
4. Simar, L.; Wilson, P.W. Estimation and inference in two-stage, semi-parametric models of production processes. *J. Econom.* **2007**, *136*, 31–64. [CrossRef]
5. Liu, J.S.; Lu, L.Y.; Lu, W.M.; Lin, B.J. A survey of DEA applications. *Omega* **2013**, *41*, 893–902. [CrossRef]
6. López-Penabad, M.-C.; Maside-Sanfiz, J.M.; Torrelles, M.J.; Iglesias-Casal, A. Application of the DEA double bootstrap to analyze efficiency in Galician sheltered workshops. *Sustainability* **2020**, *12*, 6625. [CrossRef]
7. Ngo, T.; Tsui, K.W.H. Estimating the confidence intervals for DEA efficiency scores of Asia-Pacific airlines. *Oper. Res. Int. J.* **2022**, *22*, 3411–3434. [CrossRef]
8. Stanton, K.R. Trends in relationship lending and factors affecting relationship lending efficiency. *J. Bank. Financ.* **2002**, *26*, 127–152. [CrossRef]
9. Wanke, P.; Barros, C. Two-stage DEA: An application to major Brazilian banks. *Expert Syst. Appl.* **2014**, *41*, 2337–2344. [CrossRef]
10. Yang, Z. A two-stage DEA model to evaluate the overall performance of Canadian life and health insurance companies. *Math. Comput. Model.* **2006**, *43*, 910–919. [CrossRef]
11. Byrnes, P.; Färe, R.; Grosskopf, S.; Lovell, C.A.K. The effect of unions on productivity: US surface mining of coal. *Manag. Sci.* **1988**, *34*, 1037–1053. [CrossRef]
12. Tobin, J. Estimation of relationships for limited dependent variables. *Econometrica* **1958**, *26*, 24–36. [CrossRef]
13. Hoff, A. Second stage DEA: Comparison of approaches for modelling the DEA score. *Eur. J. Oper. Res.* **2007**, *181*, 425–435. [CrossRef]
14. Kieschnick, R.; McCullough, B.D. Regression analysis of variates observed on (0, 1): Percentages, proportions and fractions. *Stat. Model.* **2003**, *3*, 193–213. [CrossRef]
15. Efron, B. Bootstrap methods: Another look at the Jackknife. *Ann. Stat.* **1979**, *7*, 1–26. [CrossRef]
16. Hirschberg, J.G.; Lloyd, P.J. *An Application of Post-DEA Bootstrap Regression Analysis to the Spill over of the Technology of Foreign-Invested Enterprises in China*; Technical Report; Department of Economics, University of Melbourne: Melbourne, Australia, 2000.
17. Simar, L.; Wilson, P.W. Statistical inference in nonparametric frontier models: Recent developments and perspectives. In *The Measurement of Productive Efficiency and Productivity Growth*; Fried, H.O., Lovell, C.A.K., Schmidt, S.S., Eds.; Oxford University Press: Oxford, UK, 2008; pp. 421–521.
18. Adler, N.; Golany, B. Including principal component weights to improve discrimination in data envelopment analysis. *J. Oper. Res. Soc.* **2002**, *53*, 985–991. [CrossRef]
19. Ferrari, S.; Cribari-Neto, F. Beta regression for modelling rates and proportions. *J. Appl. Stat.* **2004**, *31*, 799–815. [CrossRef]
20. Couri, L.; Ospina, R.; da Silva, G.; Leiva, V.; Figueroa-Zúñiga, J. A study on computational algorithms in the estimation of parameters for a class of beta regression models. *Mathematics* **2022**, *10*, 299. [CrossRef]
21. Huerta, M.; Leiva, V.; Lillo, C.; Rodriguez, M. A beta partial least squares regression model: Diagnostics and application to mining industry data. *Appl. Stoch. Model. Bus. Ind.* **2018**, *34*, 305–321. [CrossRef]

22. Davison, A.C.; Hinkley, D.V. *Bootstrap Methods and Their Application*; Cambridge University Press: Cambridge, UK, 1997.
23. Mo, R.; Huang, H.; Yang, L. An interval efficiency measurement in DEA when considering undesirable outputs. *Complexity* **2020**, *2020*, 7161628. [CrossRef]
24. Paolino, P. Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Anal.* **2001**, *9*, 325–346. [CrossRef]
25. Vasconcellos, K.L.; Cribari-Neto, F. Improved maximum likelihood estimation in a new class of beta regression models. *Braz. J. Probab. Stat.* **2005**, *19*, 13–31.
26. Figueroa-Zúñiga, J.; Niklitschek-Soto, S.; Leiva, V.; Liu, S. Modeling heavy-tailed bounded data by the trapezoidal beta distribution with applications. *Revstat-Stat. J.* **2022**, *20*, 387–404.
27. Figueroa-Zúñiga, J.; Bayes, C.L.; Leiva, V.; Liu, S. Robust beta regression modeling with errors-in-variables: A Bayesian approach and numerical applications. *Stat. Pap.* **2022**, *63*, 919–942. [CrossRef]
28. Altun, E.; El-Morshedy, M. SimBetaReg web-tool: The easiest way to implement the beta and simplex regression models. *Symmetry* **2021**, *13*, 2437. [CrossRef]
29. Buse, A. The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note. *Am. Stat.* **1982**, *36*, 153–157.
30. Nelder, J.A.; Wedderburn, R.W.M. Generalized linear models. *J. R. Stat. Soc. A* **1972**, *135*, 370–384. [CrossRef]
31. Simas, A.B.; Barreto-Souza, W.; Rocha, A.V. Improved estimators for a general class of beta regression models. *Comput. Stat. Data Anal.* **2010**, *54*, 348–366. [CrossRef]
32. Cribari-Neto, F.; Souza, T.C. Testing inference in variable dispersion beta regressions. *J. Stat. Comput. Simul.* **2012**, *82*, 1827–1843. [CrossRef]
33. White, H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econom. J. Econom. Soc.* **1980**, *48*, 817–838. [CrossRef]
34. Zeileis, A. Econometric computing with HC and HAC covariance matrix estimators. *J. Stat. Softw.* **2004**, *11*, 1–17. [CrossRef]
35. Naumann, F.; Freytag, J.; Spiliopoulou, M. Quality driven source selection using data envelope analysis. In *Proceedings of the 3rd Conference on Information Quality*; MIT Sloan School of Management: Cambridge, MA, USA, 1998; pp. 137–152.
36. Jockel, K.H. Finite sample properties and asymptotic efficiency of Monte Carlo tests. *Ann. Stat.* **1986**, *14*, 336–347. [CrossRef]
37. Lehmann, E.L.; Romano, J.P. *Testing Statistical Hypotheses*; Springer: New York, NY, USA, 2005.
38. Figueroa-Zúñiga, J.; Toledo, J.G.; Lagos-Alvarez, B.; Leiva, V.; Navarrete, J.P. Inference based on the stochastic expectation-maximization algorithm in a Kumaraswamy model with an application to COVID-19 cases in Chile. *Mathematics* **2023**, *11*, 2894. [CrossRef]
39. Ospina, R.; Ferrari, S.L.P. A general class of zero-or-one inflated beta regression models. *Comput. Stat. Data Anal.* **2012**, *56*, 1609–1623. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.