Pesquisa Operacional (2022) 42: e243504 p.1-23
doi: 10.1590/0101-7438.2022.042.00243504

Printed version ISSN 0101-7438 / Online version ISSN 1678-5142
www.scielo.br/pope
ARTICLES

ENDOGENEITY IN STOCHASTIC PRODUCTION FRONTIER WITH ONE AND TWO-STEP MODELS: AN APPLICATION WITH MUNICIPAL DATA FROM THE BRAZILIAN AGRICULTURAL CENSUS

Kessys Lorrânya Peralta de Oliveira[1*], Bernardo Borba de Andrade[2], Geraldo da Silva e Souza[2, 3] and Bruno Soares de Castro[4]

Received September 11, 2020 / Accepted January 18, 2022

ABSTRACT. Stochastic production frontier models are widely used in microeconometrics and, in the last decades, have been proven to be versatile in their range of applications. However, there are few studies concerning endogeneity in stochastic production frontier models. Here we present two stochastic production frontier models with endogenous variables based on the main distributions for the technical inefficiency. We also derive analytic gradient vectors to obtain the best performance at a reasonable computational time cost. The methodology presented here is based on one and two-step maximum likelihood estimation, allows for endogeneity and heteroscedasticity in relation to one or both error terms, and is implemented in *R* language. Finally, we illustrate an application with municipal data from the Brazilian agricultural census. The results show that capital dominates the production function, credit access and technical assistance are endogenous, and income concentration seems to impede productive inclusion through the more intensive use of technology.

Keywords: endogeneity, maximum likelihood method, stochastic production frontier.

1 INTRODUCTION

Stochastic production frontier models were simultaneously introduced by Aigner et al. (1977); Meeusen & van Den Broeck (1977) and are widely used in microeconometrics. A stochastic frontier model is a random-effects model designed to estimate the technical efficiency of decision-

*Corresponding author

[1] Federal Institute of Rondônia (IFRO), BR 435, KM 63, 76993-000, Colorado do Oeste, RO, Brazil – E-mail: kessys.oliveira@ifro.edu.br – https://orcid.org/0000-0002-9321-7756

[2] Department of Statistics, University of Brasília (UnB), Computer Science and Statistics Building, Darcy Ribeiro University Campus, 70910-900, Brasília, DF, Brazil – E-mail: bbandrade@unb.br – https://orcid.org/0000-0003-4688-9733

[3] Brazilian Agricultural Research Corporation (Embrapa/Sede), Av. W3 Norte, s/n, 70770-901, Brasília, DF, Brazil – E-mail: souzag832@gmail.com – https://orcid.org/0000-0002-6697-5383

[4] Federal University of Rondônia (UNIR), R. Rio Amazonas, 351, 76900-726, Ji-Paraná, RO, Brazil – E-mail: bruno.castro@unir.br – https://orcid.org/0000-0002-9711-3401

making units or producers through production or cost functions. The literature includes many empirical examples from various fields, such as agriculture, banking, and health.

By using cross-sectional data on the quantities of k inputs to produce a single product for each of n producers, we can write a stochastic production frontier model as

$$y_i = f(x_i;\beta)\exp(v_i)\mathscr{E}_i = f(x_i;\beta)\exp(v_i - u_i), \quad e_i = v_i - u_i, \quad i = 1,2,\ldots,n, \tag{1}$$

where y_i is the dependent variable, x_i is a vector of k inputs used by the i-th producer, β is a $k \times 1$ vector of technology parameters to be estimated, and $f(x_i;\beta)\exp(v_i)$ is the stochastic production frontier, which consists of two parts: a deterministic part, $f(x_i;\beta)$, common to all producers, and a producer-specific part, $\exp(v_i)$, which captures the effect of each producer's specific random shocks.

Since the component $\mathscr{E}_i = \exp(-u_i)$ is the production-oriented technical efficiency of the i-th producer, we have

$$\mathscr{E}_i = \frac{y_i}{f(x_i;\beta)\exp(v_i)}, \quad 0 < \mathscr{E}_i < 1, \tag{2}$$

which defines technical efficiency as the ratio between observed production and maximum feasible production, that is, $\mathscr{E}_i$ provides a measure of the observed production deficit of each producer relative to the maximum feasible production in an environment characterized by $\exp(v_i)$ (Kumbhakar & Lovell, 2003).

Estimates of each producer's technical efficiency depend on decomposing e_i and are typically derived from the conditional expectation of $\exp(-u_i)$ given e_i, which vary according to the probability density functions of both v_i and u_i.

Consider the stochastic production frontier model of (1) in the log-linear Cobb-Douglas form,

$$\ln y_i = \beta_0 + \sum_{j=1}^{k} \beta_j \ln x_{ji} + v_i - u_i. \tag{3}$$

In this specification, $\ln f(x_i;\beta)$ produces a linear model on β, with the usual Gaussian noise, $v \sim \mathbb{N}(0,\sigma_v^2)$, and a random effect, $u \sim f_u$, representing the unit's technical inefficiency.

In empirical studies, it is common to assume that u_i has a half-normal distribution. However, other distribution assumptions concerning the one-sided error term (u_i) have been proposed, such as the exponential and the truncated normal distribution. We can use maximum likelihood methods to estimate the parameters of such models. In this context, the inefficiency term is a latent variable that must be integrated when calculating the likelihood. Depending on the choice of the u_i density function, the likelihood calculation will require numerical integration. Greene (1990) and Andrade & Souza (2017) discuss approximation techniques and their accuracy. Consequently, the normal/gamma model is not available in most statistical and econometric tools for stochastic frontier analysis. Alternatively, motivated by the hierarchical structure of the stochastic frontier models, Andrade & Souza (2019) use the Expectation-Maximization (EM) algorithm to perform stochastic frontier analysis. In their work, the Expectation-Maximization calculations resulted in simple algorithms with closed-form expressions for the half-normal and exponential models and more elaborate versions for the truncated normal and gamma models.

Moreover, it is common to assume a dependency of the inefficiency component on several factors that may influence performance. With the half-normal and the exponential distributions, this is achieved by postulating a dependency of the log-likelihood on a linear construct ($\zeta' z$), defined by a set of contextual variables z. For the truncated normal, the mean of the underlying normal is defined by the linear construct. Often, we introduce heteroscedasticity in the exponential and half-normal cases, assuming that the Gaussian error terms may not have constant variance. This is achieved by also assuming that σ_v^2 is dependent on a linear function of a set of known covariates w, i.e., $\sigma_v^2 = \exp\{\rho' w\}$.

Another critical issue, as pointed out by Cazals et al. (2016), is that there may be factors (observable for the firm but unobservable for the econometrician) that affect both the choice of regressors and inefficiency levels. These factors would be endogenous variables. Consequently, failure to recognize endogeneity in the production function and the inefficiency components will lead to inconsistencies in the estimation process.

Production frontier analysis aims to identify best production practices and the importance of external factors, endogenous or not, affecting the production function and the technical efficiency component. In particular, we are interested in identifying the effect on the production of variables related to market imperfections. Following Souza et al. (2017) and Souza & Gomes (2018), market imperfections occur when farmers are subjected to different market conditions depending on their income. Large-scale farmers generally access lower input prices and sell their products at lower prices, making competition harder for small farmers. Therefore, market imperfections are typically associated with infrastructure, environment control requirements, and the presence of technical assistance. Identifying these factors and estimating the corresponding elasticities are fundamental for public policies envisaging productive inclusion.

Some programs have stochastic frontier analysis implementations available. However, often the error terms of stochastic frontier models do not have constant variance. Most stochastic frontier packages or routines available do not allow fit models considering heteroscedastic error components. For example, the *sfa* package in the *R* language permits specify half-normal, exponential, or truncated normal distributions for the one-sided error term, u. However, it does not allow you to insert a set of covariates to model the error terms. Besides that, many of them only deal with the exogeneity assumption, failing to address the endogeneity that may exist when one or more frontier or inefficiency variables correlate with the two-sided error term, v_i. Such limitations make it of interest to provide routines that enable modeling under these scenarios.

Additionally, many programs use numerical procedures to estimate parameters. However, by including the analytic gradient vector in the estimation process, we can considerably improve the convergence rate of the iterative optimization procedure. Unfortunately, in the stochastic frontier analysis literature, no work is yet available to illustrate the expressions of these gradients when endogenous variables are present.

Therefore, this paper aims to implement the prediction of the producers' technical efficiency level from stochastic production frontier models containing endogenous and exogenous variables through one and two-step maximum likelihood estimation procedure, based on Karakaplan & Kutlu (2015), considering the main specifications for the inefficiency term (half-norm, exponential, and truncated normal distributions), as well as to derive the analytic gradients of these models. Furthermore, we implemented a routine in the *R* language for these models, which makes it possible to deal with endogeneity and heteroscedasticity in any term of the model.

This article begins with a brief review of stochastic frontier models. Then, in Section 2 we cover the stochastic production frontier literature in the presence of endogeneity; Section 3 describes the database used in the application; Section 4 concerns the models and gradients derived; Section 5 discusses the results of applying these techniques to real data. Finally, Section 6 presents the conclusions reached and possible further studies.

2 LITERATURE

As the consistency of the usual stochastic frontier estimators depends on the exogeneity of the regressors, standard estimators do not deal with the endogeneity that exists if the frontier or inefficiency variables correlate with the two-sided error term, v_i, leading to inconsistent parameter estimation. Mutter et al. (2013) explains why omitting the variable causing endogeneity is not a feasible solution. Consequently, dealing with endogeneity in the stochastic frontier analysis is relatively more complicated than in standard regression models due to the unique nature of their error terms.

Some of the first stochastic frontier articles to tackle endogeneity are Guan et al. (2009); Kutlu (2010); Tran & Tsionas (2013). In these studies, variables can be endogenous because they correlate with v_i but not with u_i. Guan et al. (2009) propose a two-step estimation method to handle the endogenous regressors in the model. In the first step, they obtain consistent estimates of the frontier parameters by a generalized method of moments. In the second step, they use the residuals from the first step as a dependent variable and then use the maximum likelihood method to estimate excess capital capacity. Kutlu (2010) describes a model for dealing with endogeneity by maximum likelihood estimation at one and two-step, where he estimates the time-varying technical efficiency in the presence of endogenous regressors by using a modified version of the Battese & Coelli (1992) estimator. Tran & Tsionas (2013) propose a variation of Kutlu (2010) by a generalized method of moments. However, these model assumptions are insufficient to tackle endogeneity due to the correlation between the error terms, u_i, and v_i.

Tran & Tsionas (2015) and Amsler et al. (2016) handle endogeneity with a copula approach. Tran & Tsionas (2015) use a copula function to directly model the correlation between the endogenous regressors and the composed error term. While Amsler et al. (2016) allows the endogeneity of the regressors concerning statistical noise and inefficiency separately. A copula approach allows more general correlation structures when modeling endogeneity. However, this method is computationally intensive and requires choosing a suitable copula. In addition, the models proposed by Tran & Tsionas (2015); Amsler et al. (2016), as well as Guan et al. (2009); Kutlu (2010); Tran & Tsionas (2013) do not allow contextual variables that affect inefficiency.

Griffiths & Hajargasht (2016); Karakaplan & Kutlu (2015) handle endogeneity concerning one and two-sided errors and the correlation between them. In addition to considering environmental variables that affect inefficiency, Griffiths & Hajargasht (2016) presents a Bayesian stochastic frontier model where u_i or v_i or both correlate with the regressors. However, their model is very different from the model proposed by Karakaplan & Kutlu (2015). Instead, Karakaplan & Kutlu (2015) suggest using instrumental variables and a methodology based on a one-step maximum likelihood estimation to obtain a consistent estimator in the presence of endogeneity due to the correlation between the error terms, allowing v_i and u_i to depend on covariates that shape both distributions. In general, one of the main strengths of this model is that it is easier to apply than the copulas approach or Bayesian models, and it is a direct generalization of one of the most used stochastic frontier models - estimators of the Battese & Coelli (1995) type. In their model, Karakaplan & Kutlu (2015) assumes a linear regression by instrumental variables, but the idea can easily be generalized for a nonlinear specification. In addition, they consider a half-normal distribution for u_i. Karakaplan (2017) provides the *sfkk* module on *Stata* for this model specification.

Other recent works dealing with endogeneity are Prokhorov et al. (2020); Tsionas et al. (2021); Kumbhakar et al. (2020). Prokhorov et al. (2020) consider the problem of estimating a non-parametric stochastic frontier model with shape restrictions and when some or all regressors are endogenous. They discuss three estimation approaches based on constructing a likelihood with unknown components. Tsionas et al. (2021), using US banking data, propose a Bayesian approach for inference in the stochastic ray production frontier, which can model multiple-input - multiple-output production technologies even in case of zero output quantities. Finally, Kumbhakar et al. (2020) discuss the range of methods developed over the last four decades concerning stochastic frontier analysis.

3 DATA

For the application, we used cross-sectional data from the 2006 Brazilian agricultural census aggregated at the municipal level; from the 2010 Brazilian demographic census; from the National Institute of Research and Educational Studies (INEP), referring to education in 2009; and from the Ministry of Health in 2011. These data are valid for 4965 municipalities, which account for almost 90% of the total number of Brazilian municipalities.

The production model assumes as a dependent variable the gross income of the rural establishments in reais (*income*), i.e., the total value of the agricultural production of the establishments and, as inputs, the expenses on land (*land*), labor (*labor*) and capital (*techinputs*, technological inputs). These variables were extracted from the 2006 Brazilian agricultural census database and aggregated at the municipal level.

The contextual variables affecting production - keys to balance market imperfections - are aggregate indicators referring to the social (*social*), demographic (*demographic*) and environmental (*environment*) characteristics of the rural development. We also considered variables related to credit access (*financing*, total financing per farm), technical assistance (*techassist*, proportion of farms who received technical assistance), an indicator of income concentration per municipality (*gini*) and dummies for regional effects (*regions*, variables indicating county regions).

Except for the region, the other contextual variables were ranked and normalized by the maximum value. This approach lends nonparametric statistical properties to the analysis and circumvents outliers and heteroscedasticity problems.

For the production function, the logarithm of the gross income variable is considered a response variable. As explanatory variables we have the logarithm of the production factors - *land*, *labor* and *techinputs* - and regional dummies. The *techassist*, *financing* and *gini* variables are used to model the σ^2_{ui} function of the producers' technical inefficiency level. The *social*, *demographic* and *environment* variables are external instrumental variables in this analysis. We assumed that the *techassist* and *financing* variables are potentially endogenous. Both are complex variables that can involve many factors related to the structure of the production unit and are strong candidates for endogeneity.

4 METHODOLOGY

Consider the following stochastic production frontier model with endogenous variables proposed by Karakaplan & Kutlu (2015):

$$\begin{aligned} y_i &= x_{1i}^\top \beta + v_i - u_i, \\ x_i &= Z_i \delta + \varepsilon_i, \\ \begin{bmatrix} \tilde{\varepsilon}_i \\ v_i \end{bmatrix} &\equiv \begin{bmatrix} \Omega^{-\frac{1}{2}} \varepsilon_i \\ v_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} I_p & \sigma_{vi}\rho \\ \sigma_{vi}\rho^\top & \sigma^2_{vi} \end{bmatrix} \right). \end{aligned} \tag{4}$$

In these model, y_i is the natural logarithm of the i-th producer; x_{1i} is a vector of exogenous and endogenous variables; x_i is a $p \times 1$ vector of all endogenous variables (excluding y_i); $Z_i = I_p \otimes z_i^\top$, where z_i is a $q \times 1$ vector of all exogenous and instruments variables; v_i and ε_i are two-sided error terms; $u_i \geq 0$ is a one-sided error term capturing inefficiency; Ω is the variance-covariance matrix of ε_i; σ^2_{vi} is the variance of v_i; ρ is the vector which represents the correlation between $\tilde{\varepsilon}_i$ and v_i. In this structure, a variable is endogenous if it is not independent of the two-sided error term, v_i.

This model specifications provide a methodology for dealing with endogeneity in stochastic frontier models in a more general setting. The model considers heteroscedasticity in either component of the composed error term, allowing u_i and v_i to be dependent through covariates that shape both distributions.

Based on a Cholesky decomposition of the variance-covariance matrix of $(\tilde{\varepsilon}_i^\top, v_i)^\top$, we have:

$$\begin{bmatrix} \tilde{\varepsilon}_i \\ v_i \end{bmatrix} = \begin{bmatrix} I_p & 0 \\ \sigma_{vi}\rho^\top & \sigma_{vi}\sqrt{1-\rho^\top\rho} \end{bmatrix} \begin{bmatrix} \tilde{\varepsilon}_i \\ \tilde{w}_i \end{bmatrix}, \tag{5}$$

where $\tilde{\varepsilon}_i$ and $\tilde{w}_i \sim \mathbb{N}(0,1)$ are independent. Thus, the frontier equation is expressed by:

$$\begin{aligned} y_i &= x_{1i}^\top\beta + \sigma_{vi}\rho^\top\tilde{\varepsilon}_i + w_i - u_i, \\ &= x_{1i}^\top\beta + \frac{\sigma_{wi}}{\sigma_{cw}}\eta^\top(x_i - Z_i\delta) + e_i, \end{aligned} \tag{6}$$

where $e_i = w_i - u_i$; $w_i = \sigma_{vi}\sqrt{1-\rho^\top\rho}\tilde{w}_i = \sigma_{wi}\tilde{w}_i$; $\sigma_{cw} > 0$ is a function of the constant term of σ_{wi}; $\eta = \sigma_{cw}\Omega^{-\frac{1}{2}}\rho/\sqrt{1-\rho^\top\rho}$. Thus, when there is no heteroscedasticity in w_i, $\sigma_{wi} = \sigma_{cw}$, so that:

$$y_i = x_{1i}^\top\beta + \eta^\top(x_i - Z_i\delta) + e_i. \tag{7}$$

The term e_i is conditionally independent from the regressors given x_i and z_i, and it is possible to directly assume that the conditional distribution of v_i given x_i (and exogenous variables) is a normal distribution with mean equal to $(\sigma_{wi}/\sigma_{cw})\eta^\top(x_i - Z_i\delta)$. This approach is commonly used to solve the problem of building a consistent estimator in the presence of endogeneity for models with intrinsic nonlinearity such as this model, where $(\sigma_{wi}/\sigma_{cw})\eta^\top(x_i - Z_i\delta)$ is a bias correction term. Therefore, this approach treats endogeneity as an omitted variable problem.

Let x_{2i} be a vector of exogenous and endogenous variables and x_{3i} be a vector of exogenous and endogenous variables, which can share the same variables with x_{1i} and x_{2i}. We assume that $\sigma_{ui}^2 = \exp(x_{2i}^\top\varphi_u)$, $\sigma_{wi}^2 = \exp(x_{3i}^\top\varphi_w)$ and $\sigma_{cw}^2 = \exp(\varphi_{cw})$, where $\varphi = (\varphi_u^\top, \varphi_w^\top)^\top$ is the vector of parameters capturing heteroscedasticity, and φ_{cw} is the coefficient of the constant term for $x_{3i}^\top\varphi_w$.

The proposed specifications for the inefficiency term are: half-normal, $u_i \sim \mathbb{N}^+(0, \sigma_{ui}^2)$, exponential, $u_i \sim \text{Exp}(\sigma_{ui})$, or truncated normal, $u_i \sim \mathbb{N}^+(\mu_i, \sigma_u^2)$, with $\mu_i = x_{2i}^\top\tau$ being the mean of the truncated normal distribution, where τ is the vector of parameters capturing heteroscedasticity. For the truncated normal distribution, it is assumed that only μ_i is a function of covariates, while σ_{ui}^2 and σ_{wi}^2 are constant terms.

The log-likelihood of the stochastic production frontier model decomposes into two parts: $\ln L(\theta) = \ln L_{y|x}(\theta) + \ln L_x(\theta)$,

$$\begin{aligned}
\ln L_{y|x}(\theta) &\propto \sum_{i=1}^{n}\left\{-\frac{1}{2}\ln\sigma_i^2 + \ln\Phi\left(-\frac{e_i\lambda_i}{\sigma_i}\right) - \frac{e_i^2}{2\sigma_i^2}\right\},\\
\ln L_{y|x}(\theta) &\propto \sum_{i=1}^{n}\left\{-\frac{1}{2}\ln\sigma_{ui}^2 + \frac{\sigma_{wi}^2}{2\sigma_{ui}^2} + \ln\Phi\left(\frac{-e_i - \frac{\sigma_{wi}^2}{\sigma_{ui}}}{\sigma_{wi}}\right) + \frac{e_i}{\sigma_{ui}}\right\},\\
\ln L_{y|x}(\theta) &\propto \sum_{i=1}^{n}\left\{-\frac{1}{2}\ln\sigma^2 - \ln\Phi\left(\frac{\mu_i}{\sigma_u}\right) + \ln\Phi\left(\frac{\mu_i}{\sigma\lambda} - \frac{e_i\lambda}{\sigma}\right) - \frac{1}{2}\left(\frac{e_i+\mu_i}{\sigma}\right)^2\right\},\\
\ln L_x(\theta) &= \sum_{i=1}^{n}\left\{\frac{-p\ln(2\pi) - \ln(|\Omega|) - \varepsilon_i^\top\Omega^{-1}\varepsilon_i}{2}\right\},\\
e_i &= y_i - x_{1i}^\top\beta - \frac{\sigma_{wi}}{\sigma_{cw}}\eta^\top(x_i - Z_i\delta),\\
\varepsilon_i &= x_i - Z_i\delta,\\
\sigma_i^2 &= \sigma_{ui}^2 + \sigma_{wi}^2,\\
\lambda_i &= \frac{\sigma_{ui}}{\sigma_{wi}},\\
\mu_i &\neq 0 \iff u_i \sim \mathbb{N}^+(\mu_i, \sigma_u^2),
\end{aligned}\tag{8}$$

where $y = (y_1, \ldots, y_n)^\top$ is the vector of dependent variable, $x = (x_1^\top, \ldots, x_n^\top)^\top$ is the matrix of endogenous variables in the model and $\theta = (\beta^\top, \eta^\top, \varphi^\top, \delta^\top, \tau^\top)^\top$ is the vector of coefficients. Note that x follows a multivariate normal distribution if the number of endogenous variables is greater than one, and univariate normal otherwise. Whereas $y|x$ follows a normal/half-normal, normal/exponential or normal/truncated normal distribution, respectively. $\ln L_x(\theta)$ is added to $\ln L_{y|x}(\theta)$ and e_i is adjusted by the $(\sigma_{wi}/\sigma_{cw})\eta^\top(x_i - Z_i\delta)$ factor, which solves the problem of inconsistent parameter estimates due to endogenous regressors in x_{1i} and due to the endogenous variables in x_{2i}. Moreover, it is still possible to test for the presence of endogeneity by testing the null hypothesis that $\eta = 0$. More information in Karakaplan (2017).

In this methodology, the parameter vector θ is estimated based on a one-step maximum likelihood estimation method (simultaneously), characterizing a full information maximum likelihood (FIML) approach.

By assuming exponential link function for σ_{ui}^2 and σ_{wi}^2, and $\mu_i \in \mathbb{R}$, we obtain the following gradients:

From (8), when assuming a half-normal distribution for u_i, the gradient is

$$\begin{aligned}
U(\delta) &= \sum_{i=1}^{n} Z_i^\top \varepsilon_i \Omega^{-1} - \sum_{i=1}^{n} Z_i^\top \left\{ \frac{\sigma_{wi}}{\sigma_{cw}} \left[\frac{e_i}{\sigma_i^2} + \frac{\lambda_i}{\sigma_i} A_i \right] \right\} \eta^\top, \\
U(\beta) &= \sum_{i=1}^{n} x_{1i}^\top \left\{ \frac{e_i}{\sigma_i^2} + \frac{\lambda_i}{\sigma_i} A_i \right\}, \\
U(\eta) &= \sum_{i=1}^{n} \varepsilon_i^\top \left\{ \frac{\sigma_{wi}}{\sigma_{cw}} \left[\frac{e_i}{\sigma_i^2} + \frac{\lambda_i}{\sigma_i} A_i \right] \right\}, \\
U(\varphi_u) &= \sum_{i=1}^{n} x_{2i}^\top \left\{ \frac{1}{2\sigma_i^2} \left[\frac{e_i^2}{\sigma_i^2} - \frac{e_i}{\lambda_i \sigma_i} A_i - 1 \right] \right\} \sigma_{ui}^2, \\
U(\varphi_w) &= \sum_{i=1}^{n} x_{3i}^\top \left\{ \frac{1}{2\sigma_i^2} \left[\frac{e_i^2}{\sigma_i^2} + \frac{e_i \lambda_i}{\sigma_i} A_i \left(2 + \lambda_i^2\right) - 1 \right] \right\} \sigma_{wi}^2 + \\
&\quad + \sum_{i=1}^{n} \tilde{x}_{3i}^\top \left\{ \frac{1}{2} \eta^\top \varepsilon_i \frac{\sigma_{wi}}{\sigma_{cw}} \left[\frac{e_i}{\sigma_i^2} + \frac{\lambda_i}{\sigma_i} A_i \right] \right\},
\end{aligned} \tag{9}$$

where $A_i = \dfrac{\phi(a_i)}{\Phi(a_i)}$, $a_i = -\dfrac{e_i \lambda_i}{\sigma_i}$, and $\tilde{x}_{3i}$ is x_{3i}, except for the null-intercept component.

From (8), when assuming an exponential distribution for u_i, the gradient is

$$\begin{aligned}
U(\delta) &= \sum_{i=1}^{n} Z_i^\top \varepsilon_i \Omega^{-1} - \sum_{i=1}^{n} Z_i^\top \left\{ \frac{\sigma_{wi}}{\sigma_{cw}} \left[\frac{1}{\sigma_{wi}} B_i - \frac{1}{\sigma_{ui}} \right] \right\} \eta^\top, \\
U(\eta) &= \sum_{i=1}^{n} \varepsilon_i^\top \left\{ \frac{\sigma_{wi}}{\sigma_{cw}} \left[\frac{1}{\sigma_{wi}} B_i - \frac{1}{\sigma_{ui}} \right] \right\}, \\
U(\beta) &= \sum_{i=1}^{n} x_{1i}^\top \left\{ \frac{1}{\sigma_{wi}} B_i - \frac{1}{\sigma_{ui}} \right\}, \\
U(\varphi_u) &= \sum_{i=1}^{n} x_{2i}^\top \left\{ \frac{1}{2\sigma_{ui}^2} \left[\frac{\sigma_{wi}}{\sigma_{ui}} B_i - \frac{\sigma_{wi}^2}{\sigma_{ui}^2} - \frac{e_i}{\sigma_{ui}} - 1 \right] \right\} \sigma_{ui}^2, \\
U(\varphi_w) &= \sum_{i=1}^{n} x_{3i}^\top \left\{ \frac{1}{2\sigma_{ui}^2} + \frac{1}{2\sigma_{wi}} B_i \left[\frac{e_i}{\sigma_{wi}^2} - \frac{1}{\sigma_{ui}} \right] \right\} \sigma_{wi}^2 + \\
&\quad + \sum_{i=1}^{n} \tilde{x}_{3i}^\top \left\{ \frac{1}{2} \eta^\top \varepsilon_i \frac{\sigma_{wi}}{\sigma_{cw}} \left[\frac{1}{\sigma_{wi}} B_i - \frac{1}{\sigma_{ui}} \right] \right\},
\end{aligned} \tag{10}$$

where $B_i = \dfrac{\phi(b_i)}{\Phi(b_i)}$, $b_i = \dfrac{-e_i - \sigma_{wi}^2/\sigma_{ui}}{\sigma_{wi}}$, and $\tilde{x}_{3i}$ is x_{3i}, except for the null-intercept component.

From (8), when assuming a truncated normal distribution for u_i, the gradient is

$$\begin{aligned}
U(\delta) &= \sum_{i=1}^{n} Z_i^{\top}\varepsilon_i\Omega^{-1} - \sum_{i=1}^{n} Z_i^{\top}\left\{\frac{\sigma_w}{\sigma_{cw}}\left[\frac{e_i+\mu_i}{\sigma^2}+\frac{\lambda}{\sigma}D_i\right]\right\}\eta^{\top},\\
U(\eta) &= \sum_{i=1}^{n} \varepsilon_i^{\top}\left\{\frac{\sigma_w}{\sigma_{cw}}\left[\frac{e_i+\mu_i}{\sigma^2}+\frac{\lambda}{\sigma}D_i\right]\right\},\\
U(\beta) &= \sum_{i=1}^{n} x_{1i}^{\top}\left\{\frac{e_i+\mu_i}{\sigma^2}+\frac{\lambda}{\sigma}D_i\right\},\\
U(\tau) &= \sum_{i=1}^{n} x_{2i}^{\top}\left\{\frac{1}{\lambda\sigma}D_i - \frac{\sqrt{\lambda^{-2}+1}}{\sigma}C_i - \frac{e_i+\mu_i}{\sigma^2}\right\},\\
U(\varphi_u) &= \sum_{i=1}^{n}\left\{\frac{1}{2\sigma_u^2}\frac{\mu_i}{\sigma_u}C_i + \frac{1}{2\sigma^2}\left[\frac{(e_i+\mu_i)^2}{\sigma^2} - \frac{1}{\lambda\sigma}D_i\left(2\mu_i+e_i+\frac{\mu_i}{\lambda^2}\right)-1\right]\right\}\sigma_u^2,\\
U(\varphi_w) &= \sum_{i=1}^{n}\left\{\frac{1}{2\sigma^2}\left[\frac{(e_i+\mu_i)^2}{\sigma^2}+\frac{\lambda}{\sigma}D_i\left(\mu_i+2e_i+e_i\lambda^2\right)-1\right]\right\}\sigma_w^2,
\end{aligned} \tag{11}$$

where $C_i = \dfrac{\phi(c_i)}{\Phi(c_i)}$, $c_i = \dfrac{\mu_i}{\sigma_u}$, $D_i = \dfrac{\phi(d_i)}{\Phi(d_i)}$, and $d_i = \dfrac{\mu_i}{\sigma\lambda} - \dfrac{e_i\lambda}{\sigma}$.

Alternatively, for computationally difficult cases, as suggested by Kutlu (2010); Karakaplan & Kutlu (2015); Amsler et al. (2016), it is possible to use a two-step maximum likelihood estimation method as in Murphy & Topel (2002). In this methodology, the parameter vector θ is estimated based on a two-step maximum likelihood estimation method (separately), characterizing a limited information maximum likelihood (LIML) approach.

Besides being easier to implement, the two-step estimation process can be extended to accommodate linear or non-linear regression models by instrumental variables. Thus, in the first stage, $\ln L_x(\theta) = \ln L_1(\theta_1)$ is maximized in relation to its relevant parameters. In the second stage, conditional on the parameters estimated in the first stage, $\ln L_{y|x}(\theta) = \ln L_2(\theta_2|\theta_1)$ is maximized.

The model of the second stage is

$$y_i = x_{1i}^{\top}\beta + \eta^{\top}\hat{\varepsilon}_i + e_i, \tag{12}$$

where $e_i = w_i - u_i$ and $\hat{\varepsilon}_i$ are the estimates of the first stage residuals obtained from the equation $\hat{\varepsilon}_i = x_i - Z_i\hat{\delta}$ using ordinary least squares. Moreover, we can test the coefficients of the terms $\hat{\varepsilon}_i$ for the presence of endogeneity by testing the null hypothesis that $\eta = 0$. In this structure, a variable is endogenous if it is not independent of v_i.

Therefore, $e_i = y_i - x_{1i}^{\top}\beta - \eta^{\top}\hat{\varepsilon}_i$ and the other components are expressed in (8). As in the model proposed by Karakaplan & Kutlu (2015), in this approach, x follows a multivariate normal distribution if the number of endogenous variables is greater than one, and univariate normal otherwise. Whereas $y|x$ is specified as normal/half-normal, normal/exponential or normal/truncated normal distribution.

A disadvantage compared to the one-step procedure is that although the two-step estimation leads to consistent estimation of θ_2, the variance-covariance matrix estimated for $y|x$ needs adjust. Due to the variability in $\hat{\theta}_1$, since $\hat{\theta}_1$ is an estimate of θ_1 rather than its actual value. However, this approach presents fewer convergence problems.

Consequently, the two-step estimator provides incorrect and inconsistent standard errors, and a correction of these errors is required. To this end, an analytical approach is possible, as proposed by Murphy & Topel (2002). If the standard regularity conditions are met for both functions, then the two-step maximum likelihood estimator of θ_2 is consistent and asymptotically normally distributed with a variance-covariance matrix

$$\mathbf{V_2^*} = \mathbf{V_2} + \mathbf{V_2}(\mathbf{C}\mathbf{V_1}\mathbf{C}^\top - \mathbf{R}\mathbf{V_1}\mathbf{C}^\top - \mathbf{C}\mathbf{V_1}\mathbf{R}^\top)\mathbf{V_2}, \tag{13}$$

where

$$\begin{aligned} \mathbf{V_1} &= (q \times q) \text{ Asymptotic variance matrix of } \hat{\theta}_1 \text{ based on } \ln L_1(\theta_1), \\ \mathbf{V_2} &= (p \times p) \text{ Asymptotic variance matrix of } \hat{\theta}_2 \text{ based on } \ln L_2(\theta_2|\theta_1), \\ \mathbf{C} &= (p \times q) \text{ matrix given by } E\left[\left(\frac{\partial \ln L_2}{\partial \theta_2}\right)\left(\frac{\partial \ln L_2}{\partial \theta_1^\top}\right)\right], \\ \mathbf{R} &= (p \times q) \text{ matrix given by } E\left[\left(\frac{\partial \ln L_2}{\partial \theta_2}\right)\left(\frac{\partial \ln L_1}{\partial \theta_1^\top}\right)\right]. \end{aligned} \tag{14}$$

The matrices $\mathbf{V_1}$ and $\mathbf{V_2}$ are estimated by the respective uncorrected variance-covariance matrices, typically by the inverse matrices of negative second derivatives. At the same time, the matrices $\mathbf{C}$ and $\mathbf{R}$ are estimated by summing the individual observations on the cross products of the derivatives.

A log-likelihood is assumed to exist for the first model, $\ln L_1(\theta_1)$, as well as a conditional log-likelihood for the second (primary) model of interest, namely $\ln L_2(\theta_2|\theta_1)$. The component matrices of the Murphy-Topel estimator are estimated by the evaluation of the formulas in the maximum likelihood estimates for $\hat{\theta}_1$ and $\hat{\theta}_2$. As such,

$$\hat{\mathbf{V}}_\mathbf{2}^* = \frac{1}{n}\left[\hat{\mathbf{V}}_\mathbf{2} + \hat{\mathbf{V}}_\mathbf{2}(\hat{\mathbf{C}}\hat{\mathbf{V}}_\mathbf{1}\hat{\mathbf{C}}^\top - \hat{\mathbf{R}}\hat{\mathbf{V}}_\mathbf{1}\hat{\mathbf{C}}^\top - \hat{\mathbf{C}}\hat{\mathbf{V}}_\mathbf{1}\hat{\mathbf{R}}^\top)\hat{\mathbf{V}}_\mathbf{2}\right], \tag{15}$$

wherein

$$\begin{aligned} \hat{\mathbf{V}}_\mathbf{1} &= \left[-\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial^2 \ln f_{1i}}{\partial \hat{\theta}_1 \partial \hat{\theta}_1^\top}\right)\right]^{-1}, & \hat{\mathbf{C}} &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{2i}}{\partial \hat{\theta}_2}\right)\left(\frac{\partial \ln f_{2i}}{\partial \hat{\theta}_1^\top}\right), \\ \hat{\mathbf{V}}_\mathbf{2} &= \left[-\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial^2 \ln f_{2i}}{\partial \hat{\theta}_2 \partial \hat{\theta}_2^\top}\right)\right]^{-1}, & \hat{\mathbf{R}} &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{2i}}{\partial \hat{\theta}_2}\right)\left(\frac{\partial \ln f_{1i}}{\partial \hat{\theta}_1^\top}\right). \end{aligned} \tag{16}$$

By assuming exponential link function for σ_{ui}^2 and σ_{wi}^2, and $\mu_i \in \mathbb{R}$, we obtain the following gradients:

The gradient of the two-step model, when assuming a half-normal distribution for u_i, is

$$\begin{aligned}
U_1(\delta) &= -2\sum_{i=1}^{n} Z_i^\top \varepsilon_i,\\
U_2(\delta) &= -\sum_{i=1}^{n} Z_i^\top \left\{ \frac{e_i}{\sigma_i^2} + \frac{\lambda_i}{\sigma_i} A_i \right\} \eta^\top,\\
U_2(\eta) &= \sum_{i=1}^{n} \hat{\varepsilon}_i^\top \left\{ \frac{e_i}{\sigma_i^2} + \frac{\lambda_i}{\sigma_i} A_i \right\},\\
U_2(\beta) &= \sum_{i=1}^{n} x_{1i}^\top \left\{ \frac{e_i}{\sigma_i^2} + \frac{\lambda_i}{\sigma_i} A_i \right\},\\
U_2(\varphi_u) &= \sum_{i=1}^{n} x_{2i}^\top \left\{ \frac{1}{2\sigma_i^2} \left[\frac{e_i^2}{\sigma_i^2} - \frac{e_i}{\lambda_i \sigma_i} A_i - 1 \right] \right\} \sigma_{ui}^2,\\
U_2(\varphi_w) &= \sum_{i=1}^{n} x_{3i}^\top \left\{ \frac{1}{2\sigma_i^2} \left[\frac{e_i^2}{\sigma_i^2} + \frac{e_i \lambda_i}{\sigma_i} (2+\lambda_i^2) A_i - 1 \right] \right\} \sigma_{wi}^2,
\end{aligned} \tag{17}$$

where $A_i = \dfrac{\phi(a_i)}{\Phi(a_i)}$ and $a_i = -\dfrac{e_i \lambda_i}{\sigma_i}$.

The gradient of the two-step model assuming an exponential distribution for u_i is

$$\begin{aligned}
U_1(\delta) &= -2\sum_{i=1}^{n} Z_i^\top \varepsilon_i,\\
U_2(\delta) &= -\sum_{i=1}^{n} Z_i^\top \left\{ \frac{1}{\sigma_{wi}} B_i - \frac{1}{\sigma_{ui}} \right\} \eta^\top,\\
U_2(\eta) &= \sum_{i=1}^{n} \hat{\varepsilon}_i^\top \left\{ \frac{1}{\sigma_{wi}} B_i - \frac{1}{\sigma_{ui}} \right\},\\
U_2(\beta) &= \sum_{i=1}^{n} x_{1i}^\top \left\{ \frac{1}{\sigma_{wi}} B_i - \frac{1}{\sigma_{ui}} \right\},\\
U_2(\varphi_u) &= \sum_{i=1}^{n} x_{2i}^\top \left\{ \frac{1}{2\sigma_{ui}^2} \left[\frac{\sigma_{wi}}{\sigma_{ui}} B_i - \frac{\sigma_{wi}^2}{\sigma_{ui}^2} - \frac{e_i}{\sigma_{ui}} - 1 \right] \right\} \sigma_{ui}^2,,\\
U_2(\varphi_w) &= \sum_{i=1}^{n} x_{3i}^\top \left\{ \frac{1}{2\sigma_{ui}^2} + \frac{1}{2\sigma_{wi}} \left[\frac{e_i}{\sigma_{wi}^2} - \frac{1}{\sigma_{ui}} \right] B_i \right\} \sigma_{wi}^2,
\end{aligned} \tag{18}$$

where $B_i = \dfrac{\phi(b_i)}{\Phi(b_i)}$ and $b_i = \dfrac{-e_i - \sigma_{wi}^2/\sigma_{ui}}{\sigma_{wi}}$.

The gradient of the two-step model assuming a truncated normal distribution for u_i is

$$\begin{aligned}
U_1(\delta) &= -2\sum_{i=1}^{n} Z_i^{\top}\varepsilon_i,\\
U_2(\delta) &= -\sum_{i=1}^{n} Z_i^{\top}\left\{\frac{e_i+\mu_i}{\sigma^2}+\frac{\lambda}{\sigma}D_i\right\}\eta^{\top},\\
U_2(\eta) &= \sum_{i=1}^{n} \hat{\varepsilon}_i^{\top}\left\{\frac{e_i+\mu_i}{\sigma^2}+\frac{\lambda}{\sigma}D_i\right\},\\
U_2(\beta) &= \sum_{i=1}^{n} x_{1i}^{\top}\left\{\frac{e_i+\mu_i}{\sigma^2}+\frac{\lambda}{\sigma}D_i\right\},\\
U_2(\tau) &= \sum_{i=1}^{n} x_{2i}^{\top}\left\{\frac{1}{\lambda\sigma}D_i-\frac{\sqrt{\lambda^{-2}+1}}{\sigma}C_i-\frac{e_i+\mu_i}{\sigma^2}\right\},\\
U_2(\varphi_u) &= \sum_{i=1}^{n}\left\{\frac{1}{2\sigma_u^2}\frac{\mu_i}{\sigma_u}C_i+\frac{1}{2\sigma^2}\left[\frac{(e_i+\mu_i)^2}{\sigma^2}-\frac{1}{\lambda\sigma}D_i\left(2\mu_i+e_i+\frac{\mu_i}{\lambda^2}\right)-1\right]\right\}\sigma_u^2,\\
U_2(\varphi_w) &= \sum_{i=1}^{n}\left\{\frac{1}{2\sigma^2}\left[\frac{(e_i+\mu_i)^2}{\sigma^2}+\frac{\lambda}{\sigma}D_i\left(\mu_i+2e_i+e_i\lambda^2\right)-1\right]\right\}\sigma_w^2,
\end{aligned} \tag{19}$$

where $C_i = \dfrac{\phi(c_i)}{\Phi(c_i)}$, $c_i = \dfrac{\mu_i}{\sigma_u}$, $D_i = \dfrac{\phi(d_i)}{\Phi(d_i)}$, and $d_i = \dfrac{\mu_i}{\sigma\lambda} - \dfrac{e_i\lambda}{\sigma}$.

After obtaining estimates of the model parameters by maximum likelihood estimation, the next step is to predict the technical efficiency of each producer, $\mathscr{E}_i = \exp(-u_i)$. A natural predictor for that amount is $\hat{\mathscr{E}}_i = \exp(-\hat{u}_i)$. However, Battese & Coelli (1988) used $f(u_i|e_i)$ to derive an alternative predictor, which was modified by Battese & Coelli (1995) to take into account the heteroscedasticity that may exist regarding the error components. This alternative predictor is

$$\hat{\mathscr{E}}_i = E\{\exp(-u_i)|e_i\} = \left\{\frac{\Phi(\mu_i^*/\sigma_i^* - \sigma_i^*)}{\Phi(\mu_i^*/\sigma_i^*)}\exp\left(\frac{1}{2}\sigma_i^{*2}-\mu_i^*\right)\right\}, \tag{20}$$

where μ_i^* and σ_i^* vary according to the specification of u_i.

For the normal/half-normal model, μ_i^* and σ_i^* are

$$\begin{aligned}
\mu_i^* &= \frac{-e_i\sigma_{ui}^2}{\sigma_i^2},\\
\sigma_i^* &= \frac{\sigma_{wi}\sigma_{ui}}{\sigma_i}.
\end{aligned} \tag{21}$$

For the normal/exponential model, μ_i^* and σ_i^* are

$$\begin{aligned}
\mu_i^* &= -e_i - \frac{\sigma_{wi}^2}{\sigma_{ui}},\\
\sigma_i^* &= \sigma_{wi}.
\end{aligned} \tag{22}$$

For the normal/truncated normal model, μ_i^* and σ^* are

$$\begin{aligned} \mu_i^* &= \frac{-e_i\sigma_u^2 + \mu_i\sigma_w^2}{\sigma^2}, \\ \sigma^* &= \frac{\sigma_w\sigma_u}{\sigma}. \end{aligned} \tag{23}$$

Battese & Coelli (1988) argue that, since the production function is usually defined by the logarithm of the production ($\ln y_i$), the technical efficiency for the i-th producer should be defined as $E\{\exp(-u_i)|e_i\}$. This predictor is optimal in terms of minimizing the mean squared prediction error.

As the data are in log terms, $\hat{\mathscr{E}}_i = E\{\exp(-u_i)|e_i\}$ is a measure of the percentage by which a unit fails to reach the frontier - the ideal production rate. Thus, the closer to one $\hat{\mathscr{E}}_i$ is, the closer the producer is to achieving optimal production, with the technology incorporated into the production function.

5 RESULTS

Following the approaches present in Section 4, we fitted six models to the data described in Section 3. Thus, we estimated the model parameters in one or two-step - full information maximum likelihood (FIML) or limited information maximum likelihood (LIML) - assuming a half-normal, exponential, or truncated normal distribution for the inefficiency term, u_i. For that, we postulated a Cobb-Douglas representation in a typical stochastic frontier approach under the endogeneity assumption of technical assistance and credit access variables.

Table 1 shows some goodness of fit measures of the one and two-step models. More specifically, it presents the log-likelihood values ($\ln L$), the Akaike information criterion (AIC), and the Bayesian information criterion (BIC) of the models, as well as the Pearson correlation, the bias, and the root mean square error (RMSE) values between the observed and estimated values of the response variable, $\ln(income)$.

Table 1 – Goodness of fit measures.

Approach	Distribution of $\boldsymbol{u}$	$\ln L$	AIC	BIC	Cor$(\boldsymbol{y},\hat{\boldsymbol{y}})$	Bias	RMSE
FIML	Half-normal	-874.8	1827.5	2081.4	0.8792	0.19	1.135
	Exponential	-1330.6	2739.3	2993.2	0.8769	0.32	1.138
	Truncated normal	-432.2	944.5	1204.9	0.7315	2.87	2.895
LIML	Half-normal	-892.1	1814.3	1911.9	0.8792	0.20	1.126
	Exponential	-1284.7	2599.3	2697.0	0.8772	0.26	1.199
	Truncated normal	-435.7	903.5	1007.6	0.7636	2.88	2.901

When considering the exponential distribution for u_i, we did not obtain convergence with the BFGS optimization method, which uses analytic gradients. Therefore, we used the Nelder-Mead method, which does not use analytical gradients and is therefore slower. Until convergence, the normal/exponential model estimates using the FIML approach required 53280 iterations, whereas the LIML approach required only 4514. In both cases, compared to the other two specifications for u_i, the convergence rate when assuming an exponential distribution is considerably lower.

As for the goodness of fit measures, the normal/truncated normal models presented the lowest AIC and BIC and highest log-likelihood values, thus being the models better fitted by these criteria. However, the normal/half-normal models fitted better when considering the criteria of greater Pearson correlation and lower bias and REQM values. Additionally, all six models presented Pearson correlations between the observed and estimated values of the assumed endogenous variables (*techassist* and *financing*) greater than 0.8. Hence, indicating suitable fit of this variables from the linear regressions by instrumental variables. Based on these results and on the principle of parsimony, we selected the normal/half-normal models with their respective parameters estimated at one or two-step to model the total rural gross income of the Brazilian municipalities.

Therefore, the normal/half-normal stochastic production frontier model fitted is

$$\begin{aligned}
\ln(income_i) &= \beta_0 + \beta_1 \ln(land_i) + \beta_2 \ln(labor_i) + \beta_3 \ln(techinputs_i) + \beta_4 region_{north_i} + \\
&+ \beta_5 region_{northeast_i} + \beta_6 region_{southeast_i} + \beta_7 region_{south_i} + v_i - u_i, \\
\ln(\sigma^2_{ui}) &= \varphi_{u0} + \varphi_{u1} techassist_i + \varphi_{u2} financing_i + \varphi_{u3} gini_i, \\
\ln(\sigma^2_{wi}) &= \varphi_{w0}, \\
techassist_i &= \delta_0 + \delta_1 \ln(land_i) + \delta_2 \ln(labor_i) + \delta_3 \ln(techinputs_i) + \delta_4 region_{north_i} + \\
&+ \delta_5 region_{northeast_i} + \delta_6 region_{southeast_i} + \delta_7 region_{south_i} + \delta_8 social_i + \\
&+ \delta_9 demographic_i + \delta_{10} environment_i + \delta_{11} gini_i + \varepsilon_{1i}, \\
financing_i &= \gamma_0 + \gamma_1 \ln(land_i) + \gamma_2 \ln(labor_i) + \gamma_3 \ln(techinputs_i) + \gamma_4 region_{north_i} + \\
&+ \gamma_5 region_{northeast_i} + \gamma_6 region_{southeast_i} + \gamma_7 region_{south_i} + \gamma_8 social_i + \\
&+ \gamma_9 demographic_i + \gamma_{10} environment_i + \gamma_{11} gini_i + \varepsilon_{2i},
\end{aligned}$$

where the Center-West region is the base level.

Tables 2 and 3 provide estimates of the normal/half-normal full information and limited information models, respectively. In the Appendix (Tables 6 and 7) are the results from the linear regressions by instrumental variables of these models.

Note that, as expected for this type of model, the components of expenditure on land, labor, and capital have significant positive effects on income (Tables 2 and 3). In addition, credit access (*financing*) and income concentration (*gini*) have significant negative effects on the σ^2_{ui} function for the technical inefficiency level of agricultural properties. Indicating that greater access to rural credit and income concentration reduces the inefficiency of agricultural properties. In contrast, technical assistance (*techassist*) does not have a significant effect at 5%. This result is due to market imperfections, here represented by income concentration, which prevents technical assistance from being significant. Note the evidence of endogeneity in both cases ($\hat{\eta}_{financing}$ and $\hat{\eta}_{techassist}$ with p-value < 0.05).

Table 2 – Full information maximum likelihood estimation of the normal/half-normal model.

Variable	Coefficient	Standard error	z	P-value	Lower limit	Upper limit
Frontier						
constant	2.477	0.125	19.852	0.000	2.233	2.722
$\ln(land)$	0.110	0.014	7.639	0.000	0.082	0.139
$\ln(labor)$	0.206	0.011	17.945	0.000	0.184	0.229
$\ln(techinputs)$	0.509	0.018	28.977	0.000	0.475	0.544
region_{north}	0.066	0.055	1.201	0.230	-0.042	0.174
$\text{region}_{northeast}$	0.129	0.053	2.425	0.015	0.025	0.233
$\text{region}_{southeast}$	0.242	0.046	5.211	0.000	0.151	0.333
region_{south}	0.335	0.048	7.003	0.000	0.241	0.429
$\mathbf{ln(\sigma_u^2)}$						
constant	5.637	0.201	28.040	0.000	5.243	6.031
techassist	0.931	0.525	1.772	0.076	-0.099	1.961
financing	-2.931	0.579	-5.061	0.000	-4.067	-1.796
gini	-9.938	0.299	-33.267	0.000	-10.523	-9.352
$\mathbf{ln(\sigma_w^2)}$						
constant	-1.078	0.023	-47.806	0.000	-1.122	-1.034
$\boldsymbol{\eta_{techassist}}$						
constant	0.890	0.079	11.194	0.000	0.734	1.045
$\boldsymbol{\eta_{financing}}$						
constant	0.200	0.081	2.484	0.013	0.042	0.358
σ_w^2	0.340	0.008	44.359	0.000	0.325	0.355

Table 3 – Limited information maximum likelihood estimation of the normal/half-normal model.

Variable	Coefficient	Standard error	z	P-value	Lower limit	Upper limit
Frontier						
constant	2.552	0.126	20.306	0.000	2.306	2.799
$\ln(land)$	0.111	0.015	7.521	0.000	0.082	0.140
$\ln(labor)$	0.202	0.012	16.813	0.000	0.178	0.226
$\ln(techinputs)$	0.507	0.018	28.133	0.000	0.472	0.543
region_{north}	0.062	0.055	1.123	0.261	-0.046	0.170
$\text{region}_{northeast}$	0.131	0.053	2.455	0.014	0.026	0.235
$\text{region}_{southeast}$	0.231	0.046	4.995	0.000	0.140	0.321
region_{south}	0.319	0.047	6.736	0.000	0.226	0.412
$\mathbf{ln(\sigma_u^2)}$						
constant	6.337	0.594	10.662	0.000	5.172	7.502
techassist	-0.182	0.578	-0.315	0.753	-1.314	0.950
financing	-2.215	0.617	-3.592	0.000	-3.423	-1.006
gini	-10.441	0.841	-12.420	0.000	-12.089	-8.794
$\mathbf{ln(\sigma_w^2)}$						
constant	-1.065	0.023	-45.429	0.000	-1.111	-1.019
$\boldsymbol{\eta_{techassist}}$						
constant	0.664	0.080	8.335	0.000	0.508	0.820
$\boldsymbol{\eta_{financing}}$						
constant	0.238	0.080	2.986	0.003	0.082	0.395
σ_w^2	0.345	0.008	43.938	0.000	0.329	0.360

It is noteworthy that Souza & Gomes (2018) achieved significance for technical assistance by removing income concentration from the analysis. They concluded that the social indicator is the key variable to reduce inefficiency and reported technical assistance as a significant part of rural extension positively affecting income. In addition, improving the social indicator will facilitate access to technical assistance, thus creating a positive synergistic effect on income, reducing income concentration.

In the present work, we found that the technical assistance indicator is relatively too low for the Northern and Northeastern regions - in general, the values are less than half of the corresponding values for the other regions. Thus, public policies should be oriented to improve this indicator, especially in these regions.

Table 4 shows the results of the Wald and likelihood ratio tests on the presence of endogeneity. Note that we rejected the null hypothesis of exogeneity in both approaches - evidence of endogeneity. Therefore, to obtain consistent parameter estimates, we need fit models that take endogeneity into account.

Table 4 – Wald and likelihood ratio tests for endogeneity.

Approach	Wald	Likelihood ratio	P-value
FIML	201	342	0.000
LIML	149	208	0.000

Table 5 summarizes the relative importance of the production factors, including returns to scale for the one and two-step models. In both cases, we get decreasing returns to scale. Furthermore, capital (technological inputs) dominates the production function, followed by labor and land, showing that capital as input has a greater influence on production, corroborating the literature.

Table 5 – Relative elasticities and returns to scale.

Production factor	Coefficient	Standard error	t	P-value	Lower limit	Upper limit
FIML						
Labor	0.25	0.01	18.30	0.00	0.22	0.28
Land	0.13	0.02	7.69	0.00	0.10	0.17
Capital	0.62	0.02	31.01	0.00	0.58	0.66
Returns to scale	0.83	0.01				
LIML						
Labor	0.25	0.01	18.13	0.00	0.22	0.27
Land	0.14	0.02	7.91	0.00	0.10	0.17
Capital	0.62	0.02	31.47	0.00	0.58	0.66
Returns to scale	0.82	0.01				

Figure 1 illustrates the *box plots* for the normalized classifications of the technical efficiency measurements by region ($\hat{\mathscr{E}}_i$) predicted by the normal/half-normal models using FIML and LIML approaches. We have that efficiency differs significantly by region. Note the predominance of the

Center-West region over the others, followed by the Southeast and South, and that the North and Northeast have the lowest efficiency levels.

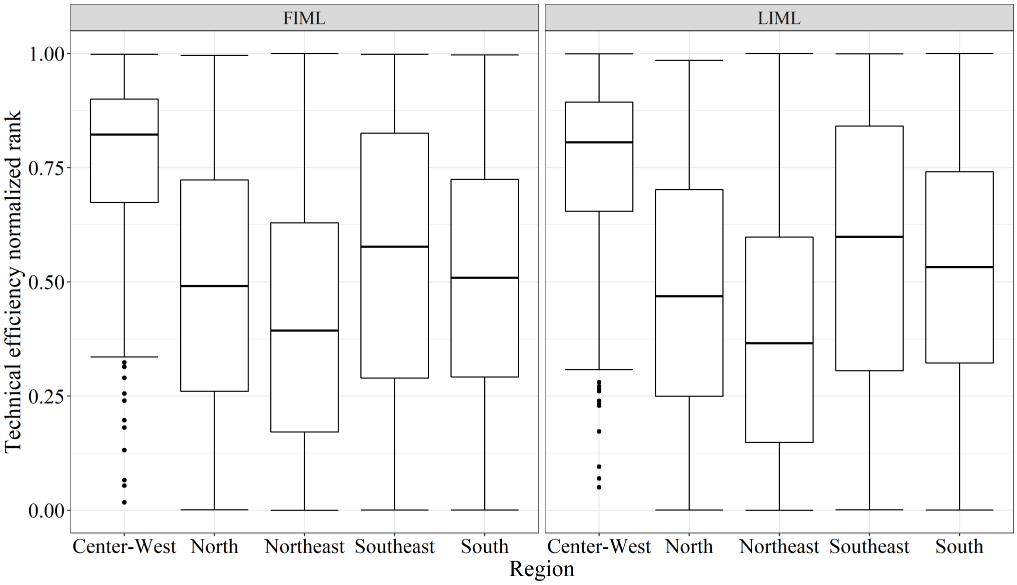

Figure 1 – *Box plots* for the technical efficiency measurements by region predicted by the normal/half-normal models using FIML and LIML approaches.

Estimates of the one and two-step procedures (FIML or LIML approaches) and predicted technical efficiencies are similar when using a normal/half-normal specification. Consequently, under standard regularity conditions, we recommend using the one-step procedure instead of the two-step one. In such conditions, the FIML estimator is more efficient than the LIML estimator and generally produces the lowest standard deviations. In contrast, we recommend using the two-step procedure for computationally intensive cases or when the one-step procedure reaches no convergence.

6 CONCLUSIONS

This paper implements the prediction of the producers' technical efficiency level from stochastic production frontier models with endogenous and exogenous variables and heteroscedastic error terms through one and two-step maximum likelihood estimation, based on Karakaplan & Kutlu (2015). We consider three main specifications for the inefficiency term (half-normal, exponential, and truncated normal distributions). We also derived the analytic gradients of these models, which are not yet available in the literature and can provide better performance at a reasonable computational time cost. Moreover, we implemented functions in the *R* language to the methodology presented here.

Additionally, we apply the models to municipal data from the Brazilian agricultural census. The application favored the use of the proposed regression models. The results from the normal/half-normal stochastic production frontier models under endogeneity are remarkably similar. Therefore, if there is convergence in the one and two-step models, then the one-step maximum likelihood estimator is recommended due to its better efficiency - smaller standard errors.

It is important to note that the correction of the variance-covariance matrix made in the two-step estimation method can change the significance of important variables compared to those in the one-step. Thus, it can change the expected technical efficiencies, especially when applying a normal/exponential model, which usually presents convergence problems.

The production function estimation is dominated by capital (technological inputs), followed by labor and land. Production shows decreasing returns to scale. Credit access and technical assistance are endogenous, and income concentration seems to impede productive inclusion through the more intensive use of technology.

Further studies are required to implement functions in *R* allowing: alternative parameterizations for the inefficiency term, such as the gamma distribution; nonlinear regressions by instrumental variables for the assumed endogenous variables; and different diagnostic analyses, such as residuals analysis. However, at the moment, a routine in the *R* language is available for the approaches described in this paper.

References

AIGNER D, LOVELL CK & SCHMIDT P. 1977. Formulation and estimation of stochastic frontier production function models. *Journal of econometrics*, **6**(1): 21–37.

AMSLER C, PROKHOROV A & SCHMIDT P. 2016. Endogeneity in stochastic frontier models. *Journal of Econometrics*, **190**(2): 280–288.

ANDRADE BB & SOUZA GS. 2017. Likelihood computation in the normal-gamma stochastic frontier model. *Computational Statistics*, pp. 1–16.

ANDRADE BB & SOUZA GS. 2019. The EM algorithm for standard stochastic frontier models. *Pesquisa Operacional*, **39**: 361–378.

BATTESE GE & COELLI TJ. 1988. Prediction of firm-level technical efficiencies with a generalized frontier production function and panel data. *Journal of econometrics*, **38**(3): 387–399.

BATTESE GE & COELLI TJ. 1992. Frontier production functions, technical efficiency and panel data: with application to paddy farmers in India. *Journal of productivity analysis*, **3**(1-2): 153–169.

BATTESE GE & COELLI TJ. 1995. A model for technical inefficiency effects in a stochastic frontier production function for panel data. *Empirical economics*, **20**(2): 325–332.

CAZALS C, FÈVE F, FLORENS JP & SIMAR L. 2016. Nonparametric instrumental variables estimation for efficiency frontier. *Journal of econometrics*, **190**(2): 349–359.

GREENE WH. 1990. A gamma-distributed stochastic frontier model. *Journal of econometrics*, **46**(1-2): 141–163.

GRIFFITHS WE & HAJARGASHT G. 2016. Some models for stochastic frontiers with endogeneity. *Journal of Econometrics*, **190**(2): 341–348.

GUAN Z, KUMBHAKAR SC, MYERS RJ & LANSINK AO. 2009. Measuring excess capital capacity in agricultural production. *American Journal of Agricultural Economics*, **91**(3): 765–776.

KARAKAPLAN MU. 2017. Fitting endogenous stochastic frontier models in Stata. *The Stata Journal*, **17**(1): 39–55.

KARAKAPLAN MU & KUTLU L. 2015. Handling endogeneity in stochastic frontier analysis. *Available at SSRN 2607276*, .

KUMBHAKAR SC & LOVELL CK. 2003. *Stochastic frontier analysis*. Cambridge university press.

KUMBHAKAR SC, PARMETER CF & ZELENYUK V. 2020. Stochastic frontier analysis: Foundations and advances I. *Handbook of production economics*, pp. 1–40.

KUTLU L. 2010. Battese-Coelli estimator with endogenous regressors. *Economics Letters*, **109**(2): 79–81.

MEEUSEN W & VAN DEN BROECK J. 1977. Efficiency estimation from Cobb-Douglas production functions with composed error. *International economic review*, pp. 435–444.

MURPHY KM & TOPEL RH. 2002. Estimation and inference in two-step econometric models. *Journal of Business & Economic Statistics*, **20**(1): 88–97.

MUTTER RL, GREENE WH, SPECTOR W, ROSKO MD & MUKAMEL DB. 2013. Investigating the impact of endogeneity on inefficiency estimates in the application of stochastic frontier analysis to nursing homes. *Journal of Productivity Analysis*, **39**(2): 101–110.

PROKHOROV A, TRAN KC & TSIONAS MG. 2020. Estimation of semi- and nonparametric stochastic frontier models with endogenous regressors. *Empirical Economics*, **60**(6): 3043–3068.

SOUZA GS & GOMES EG. 2018. A stochastic production frontier analysis of the Brazilian agriculture in the presence of an endogenous covariate. In: *International Conference on Operations Research and Enterprise Systems*. pp. 3–14. Springer.

SOUZA GS, GOMES EG & ALVES ERA. 2017. Conditional FDH efficiency to assess performance factors for Brazilian agriculture. *Pesquisa Operacional*, **37**: 93–106.

TRAN KC & TSIONAS EG. 2013. GMM estimation of stochastic frontier model with endogenous regressors. *Economics Letters*, **118**(1): 233–236.

TRAN KC & TSIONAS EG. 2015. Endogeneity in stochastic frontier models: Copula approach without external instruments. *Economics Letters*, **133**: 85–88.

TSIONAS M, IZZELDIN M, HENNINGSEN A & PARAVALOS E. 2021. Addressing endogeneity when estimating stochastic ray production frontiers: a Bayesian approach. *Empirical Economics*, pp. 1–19.

How to cite

OLIVEIRA KLP, ANDRADE BB, SILVA E SOUZA G & CASTRO BS. 2022. Endogeneity in stochastic production frontier with one and two-step models: an application with municipal data from the Brazilian agricultural census. *Pesquisa Operacional*, **42**: e243504. doi: 10.1590/0101-7438.2022.042.00243504.

APPENDIX

Table 6 – Instrumental variables regression of the one-step model.

Variable	Coefficient	Standard error	z	P-value	Lower limit	Upper limit
techassist						
constant	-0.260	0.036	-7.277	0.000	-0.331	-0.190
$\ln(land)$	0.005	0.004	1.247	0.212	-0.003	0.013
$\ln(labor)$	0.010	0.003	3.071	0.002	0.004	0.017
$\ln(techinputs)$	0.082	0.005	17.294	0.000	0.073	0.091
region_{north}	0.048	0.015	3.129	0.002	0.018	0.078
$\text{region}_{northeast}$	0.048	0.015	3.201	0.001	0.019	0.078
$\text{region}_{southeast}$	0.071	0.013	5.292	0.000	0.045	0.097
region_{south}	0.165	0.014	11.542	0.000	0.137	0.194
social	0.406	0.023	17.754	0.000	0.361	0.451
demographic	-0.016	0.028	-0.560	0.575	-0.071	0.040
environment	0.040	0.034	1.180	0.238	-0.027	0.107
gini	-0.579	0.031	-18.884	0.000	-0.639	-0.519
financing						
constant	-0.508	0.036	-13.939	0.000	-0.580	-0.437
$\ln(land)$	0.027	0.004	6.831	0.000	0.019	0.035
$\ln(labor)$	-0.005	0.003	-1.381	0.167	-0.011	0.002
$\ln(techinputs)$	0.129	0.005	26.873	0.000	0.120	0.138
region_{north}	-0.076	0.015	-4.884	0.000	-0.106	-0.045
$\text{region}_{northeast}$	-0.080	0.015	-5.241	0.000	-0.110	-0.050
$\text{region}_{southeast}$	-0.057	0.014	-4.223	0.000	-0.084	-0.031
region_{south}	0.104	0.015	7.168	0.000	0.076	0.133
social	0.156	0.024	6.590	0.000	0.109	0.202
demographic	-0.222	0.029	-7.576	0.000	-0.279	-0.165
environment	-0.398	0.036	-11.173	0.000	-0.467	-0.328
gini	-0.197	0.032	-6.087	0.000	-0.260	-0.133

Table 7 – Instrumental variables regression of the two-step model.

Variable	Coefficient	Standard error	z	P-value	Lower limit	Upper limit
techassist						
constant	-0.293	0.036	-8.148	0.000	-0.363	-0.222
$\ln(land)$	0.003	0.004	0.864	0.388	-0.004	0.011
$\ln(labor)$	0.004	0.003	1.167	0.243	-0.003	0.010
$\ln(techinputs)$	0.079	0.005	16.716	0.000	0.070	0.088
region$_{north}$	0.055	0.015	3.598	0.000	0.025	0.085
region$_{northeast}$	0.047	0.015	3.124	0.002	0.018	0.077
region$_{southeast}$	0.059	0.013	4.419	0.000	0.033	0.085
region$_{south}$	0.155	0.014	10.819	0.000	0.127	0.183
social	0.487	0.022	21.740	0.000	0.443	0.531
demographic	-0.003	0.029	-0.109	0.913	-0.060	0.054
environment	-0.018	0.035	-0.510	0.610	-0.086	0.051
gini	-0.426	0.029	-14.643	0.000	-0.483	-0.369
financing						
constant	-0.521	0.036	-14.313	0.000	-0.593	-0.450
$\ln(land)$	0.027	0.004	6.682	0.000	0.019	0.035
$\ln(labor)$	-0.007	0.003	-2.186	0.029	-0.014	-0.001
$\ln(techinputs)$	0.128	0.005	26.669	0.000	0.118	0.137
region$_{north}$	-0.073	0.015	-4.698	0.000	-0.103	-0.042
region$_{northeast}$	-0.081	0.015	-5.269	0.000	-0.111	-0.051
region$_{southeast}$	-0.062	0.014	-4.588	0.000	-0.089	-0.036
region$_{south}$	0.100	0.015	6.879	0.000	0.072	0.129
social	0.189	0.023	8.320	0.000	0.145	0.234
demographic	-0.217	0.029	-7.359	0.000	-0.275	-0.159
environment	-0.421	0.035	-11.912	0.000	-0.491	-0.352
gini	-0.134	0.030	-4.527	0.000	-0.192	-0.076