MODELLING CONDITIONAL MEASURES OF EFFICIENCY TO ASSESS THE INFLUENCE OF COVARIATES ON AGRICULTURAL PRODUCTION IN BRAZIL

Geraldo da Silva e Souza
Universidade de Brasília, Departamento de Estatística
Campus Darcy Ribeiro, Prédio CIC/EST, 70910-900, Brasília, DF, Brazil
geraldosouza@unb.br

Eliane Gonçalves Gomes
Embrapa, Superintendência de Estratégia
Parque Estação Biológica, Av. W3 Norte Final, 70770-901, Brasília, DF, Brazil
eliane.gomes@embrapa.br

ABSTRACT

Brazilian agriculture is productive, efficient, and profitable for a few farmers. However, due to the restricted access to technology, the major part of the rural producers are marginalized and outside of the mainstream of production. In this paper we assess the significance of external factors critical for productive insertion and poverty reduction in the countryside, by means of a two-stage DEA approach. The covariates considered are credit for production and exports, technical assistance, infrastructure, participation in cooperatives, education, and use of environmental friendly practices. We investigated the statistical significance of these contextual variables in production performance by means of two-stage regressions using maximum likelihood via a regional heteroskedastic beta-inflated model, which shows a superior performance over fractional regression models. The response variable is the ratio of conditional to unconditional DEA measures of efficiency, computed for each Brazilian county.

KEYWORDS. Conditional measures of efficiency. Fractional regression. Beta-inflated probability model.

Paper topics. DEA – Data Envelopment Analysis; AG&MA – OR in Agriculture and Environment.

RESUMO

A agricultura brasileira é produtiva, eficiente e lucrativa para um conjunto pequeno de agricultores. No entanto, devido ao acesso restrito a tecnologias, grande parte dos produtores rurais é marginalizada e está fora do fluxo principal da produção. Neste artigo é avaliada a importância de fatores externos para a inserção produtiva e a redução da pobreza no campo, por meio de modelagem DEA em dois estágios. As covariáveis consideradas são crédito para produção e exportação, assistência técnica, infraestrutura, participação em cooperativas, educação e uso de práticas ecologicamente corretas. A significância estatística dessas variáveis na eficiência é investigada por abordagem em dois estágios usando máxima verossimilhança e um modelo beta-inflado heterocedástico regional, que apresenta desempenho superior em relação aos modelos de regressão fracionária. A variável resposta é a razão de medidas de eficiência DEA condicionais e não condicionais, calculadas para cada município brasileiro.

PALAVRAS-CHAVE. Medidas condicionais de eficiência. Regressão fracionária. Modelo probabilístico Beta-inflacionada.

Tópicos. DEA – Análise Envoltória de Dados; AG&MA – PO na Agricultura e Meio Ambiente.

1. Introduction

Brazil is an important player in the international agricultural market, as it is the leading world exporter of several products, such as coffee, meat, soybeans, and refined sugar. According to [OECD/FAO 2015], the development of the Brazilian agricultural sector was primarily based in science and technology, which enabled the country to achieve the most advanced technology for tropical agriculture.

Recent data from the Brazilian agricultural censuses show that Brazilian agriculture is extremely income concentrated, and this evidence is persistent [Alves et al., 2020]. The 2006 agricultural census data indicated that 51% of the total production value was concentrated in 27,306 rural farms (0.62% of the total farms in Brazil) with declared annual income above 200 minimum wages [Alves et al., 2020]. The numbers in the 2017 agricultural census are not quite different. The same income class comprised about 24,000 farms, representing 0.65% of the total, and was responsible for about 53% of the total production value in 2017 [Alves et al. 2020]. The Gini index of income concentration, measured at the farm level [Souza et al. 2020a], changed from 0.85 in 2006 to 0.90 in 2017, confirming a persistent high-income concentration in agricultural production. Recent studies have pointed out that access to technology is the main causal factor of concentration and, probably, rural poverty [Souza and Gomes 2019].

Income concentration may be explained by market imperfections, resulting from different market conditions for poor and rich farmers. These may be defined [Souza and Gomes 2019] as asymmetries in access to credit for production, infrastructure, information, rural extension, and technical assistance, which inhibit the access of small farmers to technology and, therefore, to productive inclusion. On one hand, large producers can negotiate better prices for inputs and products. On the other hand, small farmers experience difficulties in adopting better technologies, as they sell their products at lower prices and buy inputs at higher prices. The evidence is that qualified labor and technological inputs are the drivers of production and productivity and that these are too expensive for rural extension to be effective in small and poor farms.

Given this background, in this paper we evaluated external factors that may affect the use of technology and lead to higher economic efficiency. We used county data derived from the 2017 Brazilian agricultural census. The modeling process updates previous studies for 2006 agricultural census [Souza et al. 2020b] and postulates that production may benefit from contextual variables reducing market imperfections. This is a two-stage analysis and follows the proposal of [Daraio and Simar 2007a, 2007b]. In the first stage, we computed a ratio of conditional to unconditional measures of efficiency, using density estimation methods in the conditioning process [Daraio and Simar 2007a, 2007b] [Bădin et al. 2012, 2014]. The ratio is used to determine whether the external factors were favorable to production in a second stage analysis. The efficiency measures computed were Free Disposal Hull – FDH and Data Envelopment Analysis – DEA, and we also investigated the effect of convexity. The statistical analysis in the second stage used a variation of the two-part fractional regression model proposed by [Ramalho et al. 2010]. Our choice to model the responses ratios was the beta-inflated model at 1 [Ospina and Ferrari 2012] and this is an extension of the proposals of [Ramalho et al. 2010] and [Souza et al. 2017, 2020b]. The major improvements are the use of DEA, the allowance of regional heteroskedasticity, and the joint maximum likelihood estimation of unit and less than unit conditional to unconditional ratio values. The method is more precise than quasi-maximum likelihood [Papke and Wooldridge 1996] and more flexible than nonlinear least squares.

2. Data

Farm-level data from the 2017 Brazilian agricultural census were aggregated by counties (municipalities). We have valid data for 5,236 counties, which represent 94% of the total municipalities in Brazil (5,570). These municipalities comprise 4,916,083 rural establishments, 97% of the total number of establishments investigated in the 2017 census (5,073,324).

Production is the gross revenue earned from agriculture (the sum of the revenues obtained from animals, vegetables, and agroindustry). The inputs are land, labor, and technology. All the variables were transformed into ranks and subsequently measured in logs. Labor was measured as the log of the municipal rank of the sum of expenditure on salary and on contracting services. Expenditures on land and technological inputs are not directly available in the census data, and we built proxies with appropriate scores. As a proxy for land, we used the log of the municipal rank of the sum of the area used in agriculture and the leased area. For technology (technological inputs), a factor model was applied to the totals of several variables. A single technological score was obtained by means of the weighted average of the ranks of these variables. The weights applied were the relative communalities derived from the factorial model. The technological inputs include: expenditures on seeds and seedlings, salt, feed, medicines, fertilizers, corrective agents, and pesticides; expenditures on fuel and electricity, other expenses, and transportation of production; capacity of warehouses, inflatables, bulk carriers, and silos; the number of tractors, seeders or planters, harvesters, fertilizers, or lime dispensers; the number of trucks, utilities, automobiles, motorcycles, airplanes, and aircraft for agricultural use; the number of animals (cattle, buffaloes, horses, donkeys, mules, pigs, goats, sheep, chickens, roosters, hens and chicks, quails, ducks, geese, drakes, partridges and pheasants, rabbits, turkeys, and ostriches); and patrimony (bee boxes, total feet of permanent crops, and total feet of forestry).

The contextual variables, tentatively identified based on previous studies [Souza and Gomes 2019] [Souza et al. 2020a, 2020b] as key factors that may affect the production process, were a municipal index of the relative use of ecological agricultural practices, financial credit, participation in cooperatives, education, infrastructure, and technical assistance. These variables are all proxies for market imperfections.

The environmentally friendly agricultural practices index is a weighted average of rankings, resulting from a factor model, with weights defined by relative communalities. The factor model is applied to the county proportions. The index includes the categories of forest management, other environmental practices (contour planting, crop rotation, soil rest, and slope conservation), and soil preparation. The financial credit index was defined as the rank obtained in the classification of the proportion of farmers who receive financial credit. The cooperative index is the rank obtained in the classification of the proportion of farmers who are members of agricultural cooperatives. The education or literacy index is the rank of the proportion of literate farmers. Infrastructure also results from a factor model applied to the ranks of the proportions of farmers with internet, telephone, and electric energy service. Technical assistance was measured by classifying the proportion of farmers who have received technical assistance.

3. Methodology

3.1. Conditional analysis

Our production model comprises n=5,236 decision making units (DMUs), one output (total rural income), and three inputs (labor, land, and technology). We computed efficiency using FDH and DEA approaches, with emphasis in the later. We assumed variable returns to scale and output orientation. Output orientation was the natural choice given that, ultimately, we seek productive inclusion and higher income gains. At this point it is worth mentioning that the FDH and DEA are not restricted to univariate production models and to output orientation [Daraio and Simar 2007a].

To consider the influence of covariates that are potentially associated with the production process, avoiding the drawbacks of [Simar and Wilson 2007], [Daraio and Simar

2007a, 2007b] proposed the analysis of conditional measurements of efficiency, assuming continuity of the covariates. These measurements are derived as follows.

Let, for each DMU τ, $X^{\tau}=(X_{\tau 1},X_{\tau 2},X_{\tau 3})$ be the vector of input variables and Y^{τ} the output variable. We assume that the production process for all DMUs is described by a joint probability measure in R^4 associated with (X,Y) [Daraio and Simar 2007b]. This measure is defined by the function $H(x,y)=\Pr\{X\le x, Y\ge y\}$ with support in the set $\Psi=\left\{(x,y)\in R^4, x=(x_1,x_2,x_3)\ge 0, \text{at least one } x_i>0, y>0 \,|\, x \text{ can produce } y\right\}$. Notice that:

1. $H(x,y)$ gives the probability that a unit operating at (input, output) levels (x,y) is dominated, i.e., that another unit produces at least as much output while using no more of any input than the unit operating at (x,y);
2. $H(x,y)$ is monotone non-decreasing in *x* and monotone non-increasing in *y*.
3. The random variables (X^{τ},Y^{τ}) are distributed as (X,Y) for all τ.

We have the decomposition shown in (1).

$$\begin{aligned} H(x,y) &= \Pr\{Y\ge y \,|\, X\le x\}\Pr\{X\le x\} \\ &= F(y\,|\,x)G(x) \end{aligned} \tag{1}$$

The output-oriented efficiency score $\theta(x,y)$, for all $(x,y)\in\Psi$, is defined by (2).

$$\begin{aligned} \theta(x,y) &= \sup_{\theta}\{H(x,\theta y)\,|\,H(x,\theta y)>0\} \\ &= \sup_{\theta}\{F(\theta y\,|\,x)\,|\,F(\theta y\,|\,x)>0, G(x)>0\} \end{aligned} \tag{2}$$

Consider production observations (x^{τ},y^{τ}), $\tau=1,\ldots,n=5{,}236$. A nonparametric estimator of the technical efficiency $\theta(x,y)$ of a producing unit (DMU), operating at (x,y), is obtained replacing $F(y\,|\,x)$ by its empirical version (3), where $I(.)$ is an indicator function.

$$\hat{F}(y\,|\,x)=\frac{\sum_{\tau=1}^{n} I(x^{\tau}\le x, y^{\tau}\ge y)}{\sum_{\tau=1}^{n} I(x^{\tau}\le x)} \tag{3}$$

Let (4) and (5). The set $\hat{\Psi}_{DEA}$ is the convex hull of $\hat{\Psi}_{FDH}$. Under convexity the two sets coincide.

$$\hat{\Psi}_{FDH}=\left\{(x,y)\in\Psi, y\le y^{\tau}, x\ge x^{\tau}, \tau=1,\cdots,n\right\} \tag{4}$$

$$\hat{\Psi}_{DEA}=\left\{(x,y)\in\Psi, y\le \textstyle\sum_{\tau=1}^{n}\gamma_{\tau}y^{\tau}, x\ge \sum_{\tau=1}^{n}\gamma_{\tau}x^{\tau}, \gamma_{\tau}\ge 0, \tau=1,\cdots,n, \sum_{\tau=1}^{n}\gamma_{\tau}=1\right\} \tag{5}$$

The estimates of the output efficiency $\theta(x,y)$ of a unit operating at (x,y) are given by (6) and (7).

$$\hat{\theta}_{FDH}(x,y)=\sup_{\theta}\left\{\theta\,|\,(x,\theta y)\in\hat{\Psi}_{FDH}\right\} \tag{6}$$

$$\hat{\theta}_{DEA}(x,y)=\sup_{\theta}\left\{\theta\,|\,(x,\theta y)\in\hat{\Psi}_{DEA}\right\} \tag{7}$$

Under free disposability $\hat{\theta}_{FDH}(x,y)$ is a consistent estimator of $\theta(x,y)$. If, additionally, Ψ is convex, $\hat{\theta}_{DEA}(x,y)$ is also consistent with a faster rate of convergence. We also have for a DMU operating at (x^{τ},y^{τ}) the definitions (8) and (9).

$$\begin{aligned}\hat{\theta}_{FDH}(x^{\tau}, y^{\tau}) &= Max\left\{\theta; \theta y^{\tau} \le \sum_{j=1}^{n} \gamma_j y^j,\ x^{\tau} \ge \sum_{j=1}^{n} \gamma_j x^j,\ \sum_{j=1}^{n} \gamma_j = 1, \gamma_j \in \{0,1\}\right\} \\ &= Max_{i:x^i \le x^{\tau}} \left\{\frac{y^i}{y^{\tau}}\right\}\end{aligned} \tag{8}$$

$$\hat{\theta}_{DEA}(x^{\tau}, y^{\tau}) = Max\left\{\theta; \theta y^{\tau} \le \sum_{j=1}^{n} \gamma_j y^j,\ x^{\tau} \ge \sum_{j=1}^{n} \gamma_j x^j,\ \sum_{j=1}^{n} \gamma_j = 1, \gamma_j \ge 0\right\} \tag{9}$$

The transformation of the data into ranks before the analysis, beyond minimizing scale issues and lending nonparametric properties to the analyses, reduces the influence of outlying observations. The underlying assumption for the DEA estimator is variable returns to scale.

The significance of a continuous contextual vector variable $Z \in \Re^{m}_{+}$ is studied considering the conditional distribution of (X,Y) given $Z = z$, which leads to the probability function $H(x, y \mid z) = \Pr\{X \le x, Y \ge y \mid Z = z\}$ and the decomposition (10).

$$\begin{aligned}H(x, y \mid z) &= \Pr\{Y \ge y \mid X \le x, Z = z\} \Pr\{X \le x \mid Z = z\} \\ &= F(y \mid x, z) G(x \mid z)\end{aligned} \tag{10}$$

Let Ψ^{z} be the support of the probability function $H(x, y \mid z)$ defined for all $(x, y) \in \Psi$, such that $G(x \mid z) > 0$. Equivalently, Ψ^{z} can be defined by the support of $F(y \mid x, z)$. The output-oriented conditional efficiency $\theta(x, y \mid z)$, given that $Z = z$, for a unit operating at (x, y) , with an environment condition $Z = z$, is defined by (11).

$$\begin{aligned}\theta(x, y \mid z) &= \sup_{\theta}\{H(x, \theta y \mid z) \mid H(x, \theta y \mid z) > 0\} \\ &= \sup_{\theta}\{F(\theta y \mid x, z) \mid F(\theta y \mid x, z) > 0, G(x \mid z) > 0\}\end{aligned} \tag{11}$$

To estimate $F(y \mid x, z)$ and to deal properly with the conditioning, let $z = (z_1, \cdots, z_m)$ and $z^{\tau} = (z^{\tau 1}, \cdots, z^{\tau m})$ be the vector of observations on Z for unit τ. [Daraio and Simar 2007a, 2007b] proposed the nonparametric density empirical estimate (12), where $K(u)$ is a multivariate kernel function with support in the unit ball $\|u\| \le 1$. The component $h_n = (h_{1n}, \cdots, h_{mn}) > 0$ is a vector of bandwidths of appropriate size.

$$\hat{F}\left(y \middle| x, z\right) = \frac{\sum_{\tau=1}^{n} I\left(x^{\tau} \le x, y^{\tau} \ge y\right) K\left(\frac{z_1 - z_1^{\tau}}{h_{1n}}, \cdots, \frac{z_m - z_m^{\tau}}{h_{mn}}\right)}{\sum_{\tau=1}^{n} I\left(x^{\tau} \le x\right) K\left(\frac{z_1 - z_1^{\tau}}{h_{1n}}, \cdots, \frac{z_m - z_m^{\tau}}{h_{mn}}\right)} \tag{12}$$

We expect that the choice of kernels and of bandwidths would not greatly influence the statistical inference process we describe later. We notice here that the estimates FDH and DEA we use below are not dependent on the kernel function, but only on the bandwidths.

[Silverman 1986] suggested the multivariate Epanechnikov model, as shown in (13), where u is a point in a *d*-dimensional space and c_d is the volume of the unit sphere. In our application, *d* = 6 and $c_6 = (1/6)\pi^3$.

$$K(u) = \begin{cases} (1/(2c_d))(d+2)(1 - u'u) & \text{if } u'u < 1 \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

The optimal bandwidth selection for the smoothing of normally distributed data with unit variance is calculated as follows. Let (14) and use $h_{in}^{opt} = v_n^{opt}, i = 1, \cdots, m$.

$$
\begin{aligned}
&v_n^{opt} = An^{-1/(d+4)} \\
&A = \left\{8c_d^{-1}(d+4)\left(2\sqrt{\pi}\right)^d\right\}^{1/(d+4)}
\end{aligned}
\tag{14}
$$

For data with no unit variances, [Silverman 1986] proposed the use of a single-scale parameter $\hat{\sigma}$ and to use the value $\hat{\sigma} v_n^{opt}$ for the window width for all components. The average variance of the covariates $\hat{\sigma}^2 = \frac{1}{d}\sum_{i=1}^{d} \hat{\sigma}_i^2$, where $(\hat{\sigma}_1^2, \cdots, \hat{\sigma}_d^2)$ is the vector of sample variances of the covariates, is an appropriate choice for the single-scale parameter $\hat{\sigma}$.

The conditional FDH and DEA empirical estimates of $\theta(x, y \mid z)$, of a unit operating at (x, y) with environmental conditions defined by z are given by (15), respectively, considering (16) and (17).

$$
\begin{aligned}
&\hat{\theta}_{FDH}(x, y \mid z) = \sup\left\{\theta \mid (x, \theta y) \in \Psi^z_{FDH}\right\}, \\
&\hat{\theta}_{DEA}(x, y \mid z) = \sup\left\{\theta \mid (x, \theta y) \in \Psi^z_{DEA}\right\}
\end{aligned}
\tag{15}
$$

$$
\hat{\Psi}^z_{FDH} = \left\{(x, y) \in \Psi \mid x \geq x^\tau, y \leq y^\tau, \text{for } \tau \text{ such that} \left|z_i - z_i^\tau\right| < h_{in}^{opt} \text{ for all } i = 1, \cdots, m\right\} \tag{16}
$$

$$
\begin{aligned}
&\hat{\Psi}^z_{DEA} = \left\{(x, y) \in \Psi \mid x \geq \sum\nolimits_{\tau \in B} \gamma_\tau x^\tau, y \leq \sum\nolimits_{\tau \in B} \gamma_\tau y^\tau\right\} \\
&B = \left\{\tau \,\middle|\, \left|z_i - z_i^\tau\right| < h_{in}^{opt} \text{ for all } i = 1, \cdots, m, \sum\nolimits_{\tau \in B} \gamma_\tau = 1, \gamma_\tau \geq 0, \tau \in B\right\}
\end{aligned}
\tag{17}
$$

We then have:

1. The set $\hat{\Psi}^z_{DEA}$ is the convex hull of $\hat{\Psi}^z_{FDH}$.
2. $\bigcup_{\tau=1}^{n} \hat{\Psi}^{z^\tau}_{FDH} = \hat{\Psi}_{FDH}, \bigcup_{\tau=1}^{n} \hat{\Psi}^{z^\tau}_{DEA} = \hat{\Psi}_{DEA}$.
3. Under disposability of Ψ, $\hat{\theta}_{FDH}(x, y \mid z)$ is consistent for $\theta(x, y \mid z)$. If, additionally, Ψ is convex, $\hat{\theta}_{DEA}(x, y \mid z)$ is also consistent.
4. For a DMU τ operating at (x^τ, y^τ, z^τ), we have (18).

$$
\begin{aligned}
&\hat{\theta}_{FDH}(x^\tau, y^\tau \left| z^\tau\right.) = \max_{\left\{i; x^i \leq x^\tau, \left|z_1^i - z_1^\tau\right| \leq h_{1n}^{opt}, \cdots, \left|z_m^i - z_m^\tau\right| \leq h_{mn}^{opt}\right\}} \left\{\frac{y^i}{y^\tau}\right\} \\
&\hat{\theta}_{DEA}(x^\tau, y^\tau \mid z^\tau) = \max\left\{\theta \mid x \geq \sum\nolimits_{\tau \in B} \gamma_\tau x^\tau, \theta y^\tau \leq \sum\nolimits_{\nu \in B} \gamma_\nu y^\nu\right\} \\
&B = \left\{\nu \,\middle|\, \left|z_i^\nu - z_i^\tau\right| < h_{in}^{opt} \text{ for all } i = 1, \cdots, m\right\} \\
&\sum\nolimits_{\nu \in} \gamma_\nu = 1, \gamma_\nu \geq 0, \nu \in B
\end{aligned}
\tag{18}
$$

5. For a DMU τ operating at (x^τ, y^τ, z^τ), we have (19).

$$
\begin{aligned}
&\hat{\theta}_{FDH}(x^\tau, y^\tau \left| z^\tau\right.) \leq \hat{\theta}_{FDH}(x^\tau, y^\tau) \\
&\hat{\theta}_{DEA}(x^\tau, y^\tau \left| z^\tau\right.) \leq \hat{\theta}_{DEA}(x^\tau, y^\tau)
\end{aligned}
\tag{19}
$$

6. The global indicator of convexity for each DMU is defined in (20).

$$
IC_\tau = \frac{\hat{\theta}_{FDH}(x^\tau, y^\tau)}{\hat{\theta}_{DEA}(x^\tau, y^\tau)} \leq 1 \tag{20}
$$

7. The conditional indicator of convexity for each DMU is defined in (21).

$$ICZ_{\tau} = \frac{\hat{\theta}_{FDH}(x^{\tau}, y^{\tau} \mid z^{\tau})}{\hat{\theta}_{DEA}(x^{\tau}, y^{\tau} \mid z^{\tau})} \leq 1 \tag{21}$$

[Daraio and Simar 2007a] suggested a nonparametric statistical analysis using the conditional ratio (22) as the response variable to study the influence of the covariates on the efficiency measurements. In (22), $\hat{\theta}(x^{\tau}, y^{\tau} \big| z^{\tau}) = \hat{\theta}_{FDH}(x^{\tau}, y^{\tau} \big| z^{\tau})$ or $\hat{\theta}_{DEA}(x^{\tau}, y^{\tau} \big| z^{\tau})$ and $\hat{\theta}(x^{\tau}, y^{\tau}) = \hat{\theta}_{FDH}(x^{\tau}, y^{\tau})$ or $\hat{\theta}_{DEA}(x^{\tau}, y^{\tau})$.

$$q(x^{\tau}, y^{\tau} \big| z^{\tau}) = \frac{\hat{\theta}(x^{\tau}, y^{\tau} \big| z^{\tau})}{\hat{\theta}(x^{\tau}, y^{\tau})} \leq 1 \tag{22}$$

As a function of z, $q(x, y \big| z)$, for output-oriented models, marginally, an increasing regression in a covariate corresponds to a favorable environmental factor, which in this case acts as a sort of extra input that is freely available to increase production. [Daraio and Simar 2007a] argued that, in this context, the value of $\hat{\theta}(x, y \big| z)$ will be closer to $\hat{\theta}(x, y)$, indicating efficiency in the use of the contextual variables.

Here we used the conditional rate with DEA measurements. The convexity indicators, globally and conditionally, in our application, were superior to 94% and we did not detect any serious depart from convexity. This supported the use of DEA. Also, DEA is a less benevolent measure of efficiency and easier to interpret under scale conditions.

3.2. Modeling the conditional ratio

We began with the inflated beta at 1 distribution. Following the parameterization of [Ospina and Ferrari 2012], the beta distribution with parameters $0 < \mu < 1, \phi > 0$ has the density function (23), where $\Gamma(.)$ is the gamma function,

$$f(u, \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} u^{\mu\phi - 1}(1-u)^{(1-\mu)\phi - 1},\ u \in (0,1) \tag{23}$$

The inflated beta at 1 distribution [Ospina and Ferrari 2012] has the density function shown in (24), where α is the probability mass at 1.

$$bi(u, \alpha, \mu, \phi) = \begin{cases} \alpha & u = 1 \\ (1-\alpha) f(u, \mu, \phi) & u \in (0,1) \end{cases} \tag{24}$$

Its mean and variance are given, respectively, by (25).

$$\begin{gathered} E(u) = \alpha + (1-\alpha)\mu \\ Var(u) = (1-\mu)\frac{\mu(1-\mu)}{\phi + 1} + \alpha(1-\alpha)(1-\mu)^2 \end{gathered} \tag{25}$$

It follows that the mean is a weighted average of the mean of a degenerate distribution and the mean of a beta distribution with weights α and $1-\alpha$. Notice that $E(u \mid u \in (0,1)) = \mu$ and $Var(u \mid u \in (0,1)) = \mu(1-\mu)/(1+\phi)$. We can see that the larger the parameter ϕ, the smaller the variance of y.

We then combined the inflated beta at 1 distribution with the two-part model proposed by [Ramalho et al. 2010] to model the conditional ratio measurements $u(z^{\tau}) = q(x^{\tau}, y^{\tau} \big| z^{\tau})$. We assumed the ratios to follow inflated beta at 1 distributions, with $\mu_{\tau} = G(\beta' z^{\tau})$, where $G(.)$ is the standard normal distribution function. We interpreted the component α as the probability of efficient use of the contextual variables. This component may be dependent of external variables.

The model can be made heteroskedastic also imposing a dependence of ϕ on a set of external factors.

The advantage of this formulation is the joint estimation of models for efficient and inefficient units in the use of covariates, by maximum likelihood, and using the complete sample. [Ramalho et al. 2010] pointed to the common use of the beta family in the case of DEA responses and, in this instance, suggested the use of the logistic, normal, or extreme value distributions for $G(.)$.

Another model allowing the use of nonlinear regression and the general method of moments (GMM), when fitting efficiency measurements, somehow robust to the beta distribution specification, follows from the expression (26) for the expected value of $u(z^\tau)$. This later formulation has been used by [Souza et al. 2017, 2020b].

$$\begin{aligned} E(u(z^\tau)) &= \alpha + (1-\alpha))E(u(z^\tau) \mid 0 < u(z^\tau) < 1) \\ E(u(z^\tau)) &= \alpha + (1-\alpha)G(\beta' z^\tau) \end{aligned} \tag{26}$$

The expectation formula is useful for computing the marginal effect of a covariate z_i, as stated in (27).

$$\frac{\partial E(u(z^\tau))}{\partial z_i} = \beta_i (1-\alpha) \frac{d}{du}\left[G'(\beta' z^\tau)\right] \tag{27}$$

4. Results and Discussion

DEA efficiency, conditional DEA, and the conditional ratio $q(x^\tau, y^\tau \mid z^\tau) = \dfrac{\hat{\theta}_{DEA}(x^\tau, y^\tau \mid z^\tau)}{\hat{\theta}_{DEA}(x^\tau, y^\tau)}$ were computed as previously described. These measures vary per county and their distribution is presented by region in Table 1, considering the output-oriented scores in the interval [1, +∞). From this table we may observe that the Northeastern region is the least efficient region. This is the region in which it is likely that the covariates will influence most. The Center-Wester region performs better.

Table 2 shows the fit resulting from the modeling of the ratio of conditional to unconditional DEA measures of efficiency using the beta-inflated probability model, as proposed in section 3.2. The rank correlation between observed and predicted values is 0.7183. All the covariates are positively associated with the ratio scores, which mean that, when it is increasing, a covariate is favorable for production. The exception is literacy. All the covariates are statistically significant. For the unit ratios, only environmental practices are not significant, and for less than unit ratios, credit is not significant. The variance components for the North and for the Center-West do not differ statistically. The ratios are less variable in the Northeastern, Southeastern, and Southern regions.

The total elasticities, as well as the marginal elasticities, vary by municipality. The medians of the total elasticity and the marginal relative elasticities for the country and regions are given in Table 3, for each covariate.

Table 1: Regional descriptive statistics for DEA efficiency, conditional DEA efficiency, and the conditional ratio.

Region	Scores	Minimum	1st Quartile	Median	3rd Quartile	Maximum
Center-West	DEA	1.000	1.060	1.087	1.116	1.228
	Conditional DEA	1.000	1.046	1.074	1.106	1.221
	Conditional ratio	0.901	0.985	0.994	0.997	1.000
North	DEA	1.000	1.086	1.108	1.133	1.303
	Conditional DEA	1.000	1.025	1.057	1.087	1.201
	Conditional ratio	0.818	0.931	0.957	0.979	1.000
Northeast	DEA	1.000	1.132	1.162	1.190	1.313
	Conditional DEA	1.000	1.036	1.069	1.102	1.240
	Conditional ratio	0.771	0.901	0.923	0.950	1.000
South	DEA	1.000	1.092	1.107	1.124	1.295
	Conditional DEA	1.000	1.070	1.090	1.108	1.254
	Conditional ratio	0.860	0.977	0.988	0.994	1.000
Southeast	DEA	1.000	1.099	1.127	1.158	1.455
	Conditional DEA	1.000	1.076	1.103	1.134	1.343
	Conditional ratio	0.819	0.971	0.991	0.996	1.000

Table 2: Maximum likelihood estimation of the ratio of conditional to unconditional measures of efficiency modeled by the beta-inflated probability model. The function $G(.)$ is the standard normal distribution function. The vector of covariates is z = (credit, cooperatives, literacy, technical assistance, environmental practices, infrastructure). D is the vector North, Northeast, Southeast, and South of dummy variables for the respective regions.

Parameter	Estimate	Standard error	p-value
$\alpha = G(b_0)$	0.0018	0.0006	<0.0001
b_0	-2.9181	0.1038	<0.0001
$\mu = G(b'z)$			
Constant	-0.8816	0.0343	<0.0001
Credit	0.0174	0.0071	0.0143
Cooperatives	0.2169	0.0045	<0.0001
Literacy	-0.2556	0.0097	<0.0001
Technical assistance	0.0820	0.0079	<0.0001
Environmental practices	0.1714	0.0077	<0.0001
Infrastructure	0.1321	0.0056	<0.0001
$\phi = \exp(d'D)$			
d_0	3.8128	0.0636	<0.0001
North	0.1135	0.0913	0.2139
Northeast	0.3414	0.0754	<0.0001
Southeast	0.3546	0.0665	<0.0001
South	0.7771	0.0696	<0.0001

We can see from Table 3 that the highest expected gains in support of production via unit relative increases in external factors will occur in the Northern and Northeastern regions. The key factors in this regard will be cooperatives, literacy, technical assistance, and environmental practices, the latter for less than unit ratios and probably inefficient counties. In this regard, we see a strong association of technology with environmental practices. Technology is the driver of production and productivity increase. Credit and technical assistance are clearly dominated by the other covariates. Credit and technical assistance are intuitively associated with higher performance scores. Favoring access to credit and to technical assistance, they should be part of agricultural public policies oriented toward alleviating poverty [World Bank 2003]. However, for this higher performance to be effective, the obstacles of the type of market imperfections to non-included farmers should be substantially reduced.

The results in Table 3 show that the influence of technical assistance is greater than that of credit. This means, considering Table 2, that rural extension is reaching out where market imperfections are relatively controlled (unit ratios). Regarding the context of Brazilian agriculture, [OECD/FAO 2015] stated that agricultural credit is the main instrument to support Brazilian productive farmers, being provided to both commercial and small-scale family farms.

Table 3: Medians of the total elasticity and of the marginal relative elasticities for Brazil and regions for each covariate.

Region	Total	Credit	Cooperatives	Literacy	Technical assistance	Infrastructure	Environmental practices
Brazil	0.0311	0.0430	0.5376	-0.6335	0.2032	0.4248	0.4248
Center-West	0.0216	0.0431	0.5376	-0.6335	0.2032	0.4248	0.4248
North	0.0424	0.0431	0.5376	-0.6335	0.2032	0.4248	0.4248
Northeast	0.1138	0.0190	0.2372	0.2794	0.0896	0.1874	0.1874
South	0.0182	0.0431	0.5376	-0.6335	0.2032	0.4248	0.4248
Southeast	0.0217	0.0431	0.5376	-0.6335	0.2032	0.4248	0.4248

We see from Table 3 that the region likely to benefit the most of an increase in all external factors is the Northeast. Apart from the Northeast, relative elasticities are about the same for all covariates.

The literacy covariate also plays an important role. In our model, it has a negative association with the performance control of market imperfections. We see this as the need to consider a higher level of education than literacy to productive inclusion. Its relative elasticity is positive only for the Northeast. Under the paradigm of emerging technologies, connectivity, and digitalization in the countryside, it is imperative to improve the qualification of the rural workforce. Farmers who are unable to use technological inputs (based on education and information) will be left out of the production process.

The effect of the covariate participation in cooperatives is positive and typically the most intense in Table 3. Cooperatives play an important role in the construction of human and social capital, with an impact on improving the managerial and organizational skills of farmers [Souza et al. 2020b]. In this sense, the agricultural performance benefits since cooperatives will help farmers to respond readily to changes in technology and market conditions.

Environmentally friendly agricultural practices are positively statistically significant with its relative effects similar to infrastructure. For units with ratio equal to one, these practices may already be part of their technological use. These practices will be transferred by technical assistance, and farmers may improve their performance thorough soil conservation and keeping forested areas, for instance. This result reinforces the idea that farms that do not use technology and preserve the environment will be outside of the mainstream of the production process.

The fact that only 0.17% of the observations are unit for the conditional ratio supports the model $E(u(z^{\tau}))=G(\beta' z^{\tau})$, which validates the assumption necessary for fractional regression. In this context, we also considered the Probit, heteroskedastic Probit (as in Stata 17) and Weibull (as in SAS 9.4) models. The later was motivated by the fact that the asymptotic distribution of the FDH estimator (which in our case is close to the DEA estimator) is in the Weibull family. The parameter estimates (contextual variables) for all fractional models point to the same direction; they show the same signs for all models, but their significances vary with the model. Table 4 shows goodness of fit measures for each model. We see that, by far, the best model is provided by the inflated beta distribution, followed by the heteroskedastic Probit, heteroskedastic Weibull, Weibull and Probit. The general formula used for fractional regression (inflated beta not included) is given by (28), where y denotes conditional DEA ratio, $G(.)$ is a distribution function, x represents the vector of covariates and $\beta, \delta, \lambda, k$ are parameters to be estimated.

$$\ln L(x, \beta, \delta, \lambda, k) = \sum_j y_j \ln\left(G(x, \beta, \delta, \lambda, k)\right) + (1 - y_j)(1 - \ln\left(G(x, \beta, \delta, \lambda, k)\right)) \quad (28)$$

The functional forms used are as follows:

1. Probit: $G(x, \beta, \delta, \lambda, k) = \Phi(x'\beta)$, where $\Phi(.)$ is the standard normal.
2. Heteroskedastic Probit: $G(x, \beta, \delta, \lambda, k) = \Phi(x'\beta / \exp(z'\delta))$, where z is a set of variables affecting the variance.
3. Weibull: $G(x, \beta, \delta, \lambda, k) = 1 - e^{-(x'\beta/\lambda)^k}$.
4. Heteroskedastic Weibull: $G(x, \beta, \delta, \lambda, k) = 1 - e^{-(x'\beta/\lambda)^k}$, $\lambda = \exp(z'\delta)$, where z is a set of variables affecting the variance.

Table 4: Goodness of fit measures for fractional regression models.

Model	-2log(likelihood)	Akaike Information
Inflated beta	-29,019.0	-28,993.0
Probit	1,556.6	1,570.6
Heteroskedastic Probit	-1,544.7	-1,566.7
Weibull	1,551.7	1,569.7
Heteroskedastic Weibull	1,541.6	1,567.6

5. Concluding Remarks

All the contextual variables are favorable to production, with exception of literacy. The median total elasticity estimated for Brazil is 0.0311. The relative elasticities are higher for participation in cooperatives (0.5376), ecological agricultural practices (0.4248), and literacy (-0.6335). The negative effect of literacy in all regions, except in the Northeast, is interpreted as the need for a variable reflecting a higher level of education to be included in the model. The factor ecological agricultural practices is heavily associated with technology, the main driver leading to productive insertion and poverty reduction. The effect of participation in a cooperative is strong for all regions.

The highest expected gains from public policies, envisaging proper control of contextual variables and reduction of market imperfections, will occur in the North and Northeast. Public policies should focus on providing fair market opportunities for small and large farmers. Access to credit and to technical assistance and the workforce qualification level restrict the adoption of innovations, with emphasis on the adoption of low-carbon agricultural production systems. When considering the new paradigm of 4.0 agriculture, these covariates are also of importance, as they are responsible for the diffusion and adoption of new technologies. These are important to strengthen and to transform Brazilian agriculture in the context of the digital agriculture and of the climate change agenda.

References

Alves, E.R.A., Souza, G.S. and Gomes, E.G. (2020). A concentração do valor bruto da produção e a pobreza segundo o Censo Agropecuário 2017. In Navarro, Z. (ed.), *A economia agropecuária do Brasil*: a grande transformação, Baraúna, 176–182.

Bădin, L., Daraio, C. and Simar, L. (2012). How to measure the impact of environmental factors in a nonparametric production model. *European Journal of Operational Research* 223: 818–33.

Bădin, L., Daraio, C. and Simar, L. (2014). Explaining inefficiency in nonparametric production models: the state of the art. *Annals of Operations Research* 214 (1): 5–30.

Conover, W. J. 1999. *Practical Nonparametric Statistics*. 3rd ed. New York: Wiley.

Daraio, C. and Simar, L. (2007a). *Advanced Robust and Nonparametric Methods in Efficiency Analysis*. Springer, New York.

Daraio, C. and Simar, L. (2007b). Conditional nonparametric frontier models for convex and nonconvex technologies: a unifying approach. *Journal of Productivity Analysis* 28: 13-32.

IBGE (2019). *Censo Agropecuário 2017*. https://sidra.ibge.gov.br/pesquisa/censo-agropecuario/censo-agropecuario-2017 Accessed 2020-03-06.

OECD/FAO (2015). *OECD–FAO Agricultural Outlook 2015*. Paris: OECD Publishing. http://dx.doi.org/10.1787/agr_outlook-2015-en Accessed 2020-03-06.

Ospina, R. and Ferrari, S.L.P. (2012). A General class of zero-or-one inflated beta regressions models. *Computational Statistics and Data Analysis* 56: 1609–23.

Papke, L.E. and Wooldridge, J.M. (1996). Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Economics* 11 (6): 619–32.

Ramalho, E.A., Ramalho, J.J.S. and Henriques, P.D. (2010). Fractional regression models for second stage DEA efficiency analyses. *Journal of Productivity Analysis* 34: 239–55.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall/CRC, Florida.

Simar, L. and Wilson, P.W. (2007). Estimation and inference in two-stage, semi-parametric models of production processes. *Journal of Econometrics* 136: 31–64.

Souza, G.S. and Gomes, E.G. (2019). A stochastic production frontier analysis of the Brazilian agriculture in the presence of an endogenous covariate. In Parlier, G., Liberatore, F. and Demange, M. (eds.), *Operations Research and Enterprise Systems—ICORES 2018. Communications in Computer and Information Science*, 966, 3–14.

Souza, G.S., Gomes, E.G. and Alves, E.R.A. (2017). Conditional FDH efficiency to assess performance factors for Brazilian agriculture. *Pesquisa Operacional* 37: 93–106.

Souza, G.S., Gomes, E.G. and Alves, E.R.A. (2020a). Estimativa de uma função de produção para a agricultura brasileira com base nos microdados do censo agropecuário de 2017. *Revista de Política Agrícola* 29:65–82.

Souza, G.S., Gomes, E. G. and Alves, E.R. (2020b). Two-part fractional regression model with conditional FDH responses: an application to Brazilian agriculture. *Annals of Operations Research*. https://doi.org/10.1007/s10479-020-03752-z.

World Bank (2003). *Rural Poverty Alleviation in Brazil: Toward an Integrated Strategy*. Clearance Center, Danvers.