MÉTODOS ALTERNATIVOS PARA ANÁLISE DE REGRESSÃO EM DOIS ESTÁGIOS COM RESPOSTA DEA: UMA APLICAÇÃO PARA A AGRICULTURA BRASILEIRA

Geraldo da Silva e Souza
Embrapa, Secretaria de Inteligência e Relações Estratégicas
PqEB, Av. W3 Norte final, 70770-901, Brasília, DF, Brasil
geraldo.souza@embrapa.br

Eliane Gonçalves Gomes
Embrapa, Secretaria de Inteligência e Relações Estratégicas
PqEB, Av. W3 Norte final, 70770-901, Brasília, DF, Brasil
eliane.gomes@embrapa.br

RESUMO

Neste artigo comparam-se métodos *bootstrap* e de regressão fracionária para a avaliação da influência de covariáveis em modelos de regressão em dois estágios com respostas DEA. O último faz uso do método de quasi-verossimilhança e regressão não linear e o primeiro faz uso de repetições *bootstrap* obtidas apenas para unidades ineficientes. Sugere-se um modelo para a inclusão de variáveis endógenas na análise com base no modelo *two-part*, com modelagem distinta para unidades eficientes e ineficientes. A aplicação em apreço estuda o efeito das variáveis contextuais crédito, assistência técnica, indicador social, indicador ambiental, indicador demográfico, concentração de renda e variáveis indicadoras regionais na eficiência de produção agrícola municipal. A eficiência é calculada segundo o modelo DEA-VRS com orientação a *output*. Essa hipótese é validada via *bootstrap*. Em geral, os modelos apresentam resultados inferenciais diferentes. Conclui-se do modelo *two-part* que a assistência técnica é positiva para a produção. A região Centro-oeste possui maior probabilidade de municípios, na mediana, tornaram-se eficientes, seguindo-se as regiões Sul, Sudeste, Norte e Nordeste, nesta ordem.

Palavras-chave: DEA; Variáveis contextuais; *Bootstrap*; Regressão fracionária; Agricultura.

ABSTRACT

In this paper we compare bootstrap and fractional regression methods for the evaluation of the influence of covariates in two-stage regression models with DEA responses. The latter uses quasi-likelihood and nonlinear regression method and the former bootstrap repetitions obtained only for inefficient units. It is suggested a model for the inclusion of endogenous variables based on the two-part model, with distinct modeling for efficient and inefficient units. The application assesses the effect of the contextual variables credit, technical assistance, social indicator, environmental indicator, demographic indicator, income concentration and regional indicator variables on the efficiency of municipal agricultural production. Efficiency is calculated according to a DEA-VRS model with output

orientation. This hypothesis is validated via bootstrap. In general, the models have different inferential results. It is concluded from the two-part model that technical assistance is positive for production. The Center-west region has the highest probability of municipalities, in the median, became efficient, followed by the South, Southeast, North and Northeast, in this order.

Keywords: DEA; Contextual variables; Bootstrap; Fractional regression; Agriculture.

Como Citar:
SOUZA, Geraldo da Silva e; GOMES, Eliane Gonçalves. Métodos alternativos para análise de regressão em dois estágios com resposta DEA: uma aplicação para a agricultura brasileira . *In*: SIMPÓSIO DE PESQUISA OPERACIONAL E LOGÍSTICA DA MARINHA, 19., 2019, Rio de Janeiro, RJ. **Anais** […]. Rio de Janeiro: Centro de Análises de Sistemas Navais, 2019.

1. INTRODUÇÃO

É recorrente na literatura de análise de eficiência com hipótese determinística da fronteira de produção o interesse em avaliar o efeito de variáveis contextuais (ou covariáveis) na resposta, usualmente uma medida de desempenho do tipo DEA (*Data Envelopment Analysis*), com valores normalizados no intervalo (0,1]. A análise tipicamente é feita em dois estágios. No primeiro estágio calculam-se as medidas de eficiência DEA e, ulteriormente, acessa-se a associação dessas medidas com as variáveis contextuais.

As primeiras aplicações deste tipo utilizavam-se de métodos de regressão linear ou, dada a natureza limitada e truncada dos valores da resposta, regressões do tipo Tobit. Tais métodos foram extensivamente criticados na literatura, particularmente nos artigos de Simar e Wilson (2007, 2011). Banker e Natarajan (2008, 2011) sugerem métodos de programação linear para avaliar o impacto de variáveis contextuais na medida de eficiência. De um modo geral, a identificação de fatores causais da eficiência determinística demanda análise estatística apropriada, como alertam Souza e Gomes (2015). Nesse contexto, como apontado pelos mesmos autores, surgem dois problemas estatísticos principais: correlação entre as medidas de eficiência e potencial endogeneidade das variáveis contextuais. A correlação entre as medidas DEA é induzida pelo método de cálculo e nem sempre é problemática, dada a consistência dessas medidas sob a hipótese da existência de um modelo de produção subjacente. Em estudos de modelos de análise de variância com *input* unitário, por exemplo, o problema não existe (Gomes et al., 2008). Endogeneidade é um problema mais envolvente e invalida todas as abordagens citadas acima.

Simar e Wilson (2007), na ausência de endogeneidade, i.e., sob a hipótese de separabilidade, sugerem dois métodos de análise envolvendo a distribuição normal truncada e o uso intensivo da técnica *bootstrap*. A hipótese de não separabilidade pode ser tratada com uso do FDH condicional de Daraio e Simar (2007) e Bădin et al. (2012) via regressão não paramétrica. Nos casos em que a endogeneidade não está presente, Ramalho et al. (2010) sugerem o uso de regressões fracionárias explorando a técnica de análise da quasi-verossimilhança proposta por Papke e Wooldridge (1996). Ramalho et al. (2010) sugerem, ainda, o uso de um modelo em duas partes (*two-part model*), com o qual é possível a estimativa da probabilidade de uma DMU tornar-se e eficiente.

Neste artigo comparam-se as técnicas de regressão fracionária de Ramalho et al. (2010) e os algoritmos de Simar e Wilson (2007) na avaliação da eficiência técnica da produção agrícola municipal. No segundo estágio postula-se a medida de eficiência DEA

como dependente das variáveis contextuais crédito rural, assistência técnica, indicador social, indicador ambiental, indicador demográfico, concentração de renda e variáveis indicadoras regionais. Testa-se a hipótese de retornos constantes a partir dos resultados de Simar e Wilson (2002). Adicionalmente, sugere-se uma abordagem GMM (*Generalized Method of Moments*) para a avaliação da endogeneidade a partir do modelo *two-part*.

2. ESTUDO DE CASO

O setor agropecuário é, sem dúvida, um setor crucial para a economia brasileira. O Brasil lidera a pauta de exportações mundiais de diversos produtos e, segundo OECD/FAO (2015), pode se tornar o maior fornecedor de alimentos do mundo, capaz de satisfazer a demanda mundial adicional, proveniente principalmente da Ásia. A agricultura brasileira, contudo, é extremamente concentrada (Souza et al., 2013). Seu potencial de produção poderá ser substancialmente aumentado por meio de inclusão produtiva, tornando mais efetiva a agricultura intensiva em insumos tecnológicos, principalmente para a pequena produção incipiente e afastada dos mercados de devido a imperfeições econômicas no crédito e acesso a informação, logística, educação etc. (Souza et al., 2018). Desse modo, justificam-se estudos pertinentes à otimização de sistemas de produção e da influência de variáveis contextuais potencialmente causais de incrementos significativos na produção, via aumento de produtividade e eficiência.

Nossa aplicação neste artigo visa identificar fatores causais da eficiência técnica agrícola municipal. Para esse fim, utilizam-se os dados de produção do Censo Agropecuário de 2006. Saliente-se que não estão ainda disponíveis os (micro)dados do Censo Agropecuário de 2017 (devem ser divulgados no segundo semestre de 2019). A abordagem para o cálculo das medidas DEA é a mesma utilizada em Souza et al. (2016).

A variável de *output* é a renda bruta da produção e os *inputs* são despesas com mão de obra, com terra e com insumos tecnológicos. As variáveis contextuais consideradas no estudo são: total do crédito agrícola (*rank* normalizado pelo máximo), proporção de estabelecimentos que receberam orientação técnica, índice de concentração da renda para o município (índice de Gini), indicador ambiental, indicador social e indicador demográfico. Os três últimos indicadores são escores entre 0 e 1 e estão descritos em mais detalhes em Souza et al. (2016). A dimensão social reflete o nível de bem-estar propiciado por fatores como disponibilidade de água e energia elétrica e forma de esgotamento sanitário no domicílio rural. Reflete também indicadores do nível de educação e de saúde, diretamente relacionados com melhor qualidade de vida dos residentes, e o nível de pobreza e rendimentos mensais per capita dos domicílios rurais. A dimensão demográfica capta os aspectos relacionados à dinâmica da população, que tende a acompanhar o desenvolvimento rural. A dimensão ambiental capta aspectos relacionados a boas práticas de manejo do solo, medidas dos níveis de preservação e conservação de matas e florestas, e uso de produtos agroquímicos nos estabelecimentos rurais. Foram também consideradas variáveis binárias (*dummy*) indicativas das regiões geográficas brasileiras. As análises foram feitas em relação à região Centro-oeste e, por isto, esta região não aparece diretamente nas tabelas de resultados. Os dados de produção compreendem 4.964 municípios brasileiros com dados válidos em todas as variáveis consideradas.

3. ASPECTOS METODOLÓGICOS

3.1. DEA

Considere-se um processo de produção composto por 4.964 firmas (municípios). Cada município faz uso do vetor de insumos $(x_1, x_2, x_3)'$ – terra, trabalho e insumos

tecnológicos – para produzir o nível de produto *y*. Seja $Y=(y_1,y_2,\ldots,y_{4964})$ o vetor produto do processo. Seja *X* a matriz 3×4964 de uso dos insumos. A *r*-ésima coluna de *X* é o vetor de insumos utilizados pelo município *r*.

A medida de eficiência técnica de produção $\phi^*(x_o,y_o)$, mais geralmente de desempenho, para o município *o*, com vetor de produção (x_o,y_o), derivada da abordagem de DEA, orientada a produto e com retornos variáveis à escala (VRS) é dada por (Banker et al., 1984):

$\phi^*(x_o,y_o)=\max_{\phi,\lambda}\phi$

sujeito a

i) $Y\lambda\geq\phi y_o$, ii) $X\lambda\leq x_o$ e iii) $\lambda\geq0, \lambda1=1,\ \phi$ livre

Duas modelagens são utilizadas no artigo. Na análise de regressão fracionária, no modelo *two-part* e no primeiro algoritmo de Simar e Wilson (2007) usa-se a medida $0<(1/\phi^*)\leq1$ como resposta. No segundo algoritmo de Simar e Wilson (2007) usa-se $\phi^*\geq1.$ Esta modelagem também foi usada para o Algoritmo 1.

3.2. ABORDAGEM DE SIMAR E WILSON (2007)

Sob condições de separabilidade, Simar e Wilson (2007) propõem os dois algoritmos adaptados abaixo para $1/\phi^*$. A família de densidades paramétricas utilizada é a da normal $N(\mu,\sigma^2)$ truncada em (0,1), cuja expressão vem dada por:

$$f(\varepsilon)=\begin{cases}\frac{1}{\sigma}g(\frac{\varepsilon-\mu}{\sigma})/\left(G(\frac{1-\mu}{\sigma})-G(\frac{-\mu}{\sigma})\right), & se\ \ 0<\varepsilon<1\\ 0, & c.c.\end{cases}$$

As funções $g(.)$ e $G(.)$ representam, respectivamente, as funções de densidade e de distribuição de probabilidades da normal padrão. Postula-se que $\mu=z\beta'$, sendo z o vetor de observação nas variáveis contextuais, de dimensão p, e β um parâmetro p-dimensional desconhecido.

O vetor de parâmetros β é estimado pela maximização da função de verossimilhança conjunta das *n* medidas de eficiência observadas no intervalo (0,1):

$$L(\beta,\sigma)=\sum_{r=1}^{n}\left(-\log(\sigma)+\log(\phi\left(\tfrac{x_r-\mu_r}{\sigma}\right))-\log\left[\Phi(\tfrac{1-\mu_r}{\sigma})-\Phi(\tfrac{-\mu_r}{\sigma})\right]\right)$$

A distribuição normal truncada em (0,1) pode ser caracterizada por $\varphi_r=\mu_r+\varepsilon_r$, onde a variável aleatória ε_r é a normal $N(0,\sigma^2)$ truncada à esquerda em $-\mu_r$ e à direita em $1-\mu_r$.

3.2.1. Algoritmo 1

1. Calcule ϕ_r^* para cada município *r*.

2. Use o método de máxima verossimilhança para obter um estimador $\hat{\beta}$ de β, bem como um estimador $\hat{\sigma}$ de σ na regressão normal truncada de ϕ_r^* em $z_r\beta'$, eliminando as observações com eficiência unitária.
3. Repita os passos (3.1-3.3) abaixo L vezes para obter o conjunto de repetições *bootstrap* $A=\{(\hat{\beta}^*,\hat{\sigma}^*)_b\}_{b=1}^{L}$.
   3.1 Para $r=1,...,n$, gere ε_r da distribuição $N(0,\hat{\sigma}^2)$ truncada à esquerda em $-z_r\hat{\beta}'$ e à direita em $1-z_r\hat{\beta}'$.
   3.2 Para $r=1,...,n$ calcule $\theta_r^*=z_r\hat{\beta}'+\varepsilon_r^*$.
   3.3 Use o método de máxima verossimilhança para estimar a regressão de θ_r^* em z_r, obtendo novas estimativas $(\hat{\beta}^*,\hat{\sigma}^*)$.
4. Use as repetições *bootstrap* em A e os estimadores originais $\hat{\beta}$, $\hat{\sigma}$ para obter estimativas *bootstrap* de parâmetros, desvios padrão e intervalos de confiança.

3.2.2. Algoritmo 2

1. Calcule ϕ_r^* para cada município *r*.
2. Use o método de máxima verossimilhança para obter um estimador $\hat{\beta}$ de β, bem como um estimador $\hat{\sigma}$ de σ na regressão normal truncada de ϕ_r^* em $z_r\beta'$ eliminando as observações com eficiência unitária.
3. Repita os passos (3.1-3.4) L_1 vezes para obter os *n* conjuntos de repetições *bootstrap* $B_r=\{\hat{\theta}_{rb}^*\}_{b=1}^{L_1}$.
   3.1 Para $r=1,...,n$, gere ε_r da distribuição $N(0,\hat{\sigma}^2)$ truncada à esquerda em $-z_{rb}\hat{\beta}'$ e à direita em $1-z_{rb}\hat{\beta}'$.
   3.2 Para $r=1,...,n$. Calcule $\theta_{rb}^*=z_{rb}\hat{\beta}+\varepsilon_{rb}^*$.
   3.3 Calcule $x_{rb}^*=\left(\phi_r^*/\theta_{rb}^*\right)x_r$, para todo $r=1,...,n$. Esta correção dos insumos é necessária e é herdada de Simar e Wilson (1998).
   3.4 Calcule o estimador *bootstrap* $\hat{\theta}_{rb}^*$ de ϕ_r^*, para $r=1,...,n$, resolvendo
   $$\hat{\theta}_{rb}^*=\min\left\{\theta;\ y_r\le\sum_{i=1}^{n}\gamma_i y_i;\ \theta x_r\ge\sum_{i=1}^{n}\gamma_i x_{rb}^*;\ \theta>0;\ \gamma_i\ge 0,\ i=1,...,n\right\}.$$
4. Para $r=1,...,n$, calcule o estimador viés corrigido $\tilde{\theta}_r=\left(\frac{1}{L_1}\sum_{b=1}^{L_1}\hat{\theta}_{rb}^*\right)-\phi_r^*=\bar{\theta}_r^*-\phi_r^*$.
5. Use o método de máxima verossimilhança para estimar a regressão normal truncada de $\tilde{\theta}_r$ em z_r, obtendo as estimativas $(\tilde{\beta},\tilde{\sigma})$.
6. Repita os passos (6.1-6.3) L_2 vezes para obter um conjunto de repetições *bootstrap* $C=\{(\hat{\beta}^*,\hat{\sigma}_\varepsilon^*)_b\}_{b=1}^{L_2}$.
   6.1 Para $r=1,...,n$, gere ε_r da distribuição $N(0,\tilde{\sigma})$ truncada à esquerda em $-z_r\tilde{\beta}'$ e à direita em $1-z_r\tilde{\beta}'$.

6.2 Para $r=1,...,n$ calcule $\theta_r^{**}=z_r\tilde{\beta}'+\varepsilon_r$.

6.3 Use o método de máxima verossimilhança para estimar a regressão truncada de θ_r^{**} em z_r, obtendo as estimativas $(\hat{\beta}^*,\hat{\sigma}^*)$.

7. Use as repetições *bootstrap* em C e as estimativas originais $(\tilde{\beta},\tilde{\sigma})$ na obtenção de estimativas *bootstrap* dos parâmetros, desvios padrão e intervalos de confiança.

Observe-se que as duas abordagens *bootstrap* exigem realizações iid da distribuição $N(\mu,\sigma^2)$ truncada em (0,1). Conhecidos μ e σ, essas realizações são geradas utilizando a expressão seguinte, onde w é uma variável aleatória com distribuição uniforme em (0,1).

$$t=\mu+\sigma G^{-1}\left[wG\left(\frac{1-\mu}{\sigma}\right)+G\left(\frac{-\mu}{\sigma}\right)(1-w)\right] .$$

3.3. REGRESSÃO FRACIONÁRIA E O MODELO *TWO-PART*

Os modelos de regressão fracionária que serão aqui usados estão descritos em detalhes em Ramalho et al. (2010). Para avaliar estatisticamente o efeito de uma covariável na resposta $0<\phi^*\leq 1$, postula-se que $E(\phi^*\mid z)=F\left(z\beta'\right)$, onde $F(.)$ é uma função de distribuição de probabilidades. O modelo é viável mesmo na presença de observações com eficiência unitária. O parâmetro β é estimado pelo método de quasi-verossimilhança, por meio da maximização da função

$$\sum_{r=1}^{4964}\left(\phi_r^*\log\left(F\left(z_r\beta'\right)\right)+(1-\phi_r^*)\log\left(1-F\left(z_r\beta'\right)\right)\right)$$

Se a média for especificada corretamente, independentemente da distribuição condicional verdadeira, $\sqrt{n}\left(\hat{\beta}-\beta\right)\xrightarrow{d}N(0,V)$. A matriz V é estimada conforme abaixo. O estimador de quasi-verossimilhança é eficiente para certa classe de estimadores que contém a quasi-verossimilhança para a família exponencial e mínimos quadrados não lineares ponderados.

$$\hat{V}=\left(\hat{A}\right)^{-1}\hat{B}\hat{A},\quad \hat{A}=1/4964\sum_{r=1}^{4964}\left(\hat{f}_r^2/\hat{F}_r(1-\hat{F}_r)\right)z_r{}'z_r$$
$$\hat{B}=1/4964\sum_{r=1}^{4964}\left(\hat{u}_r^2\hat{f}_r^2/(\hat{F}_r(1-\hat{F}_r))^2\right)z_r{}'z_r$$
$$\hat{F}_i=F(z_r\hat{\beta}'),\quad \hat{f}_r=F'(z_r\hat{\beta}'),\quad \hat{u}_r=\phi_r^*-\hat{F}_r$$

É claro que medidas mais robustas para os desvios dos parâmetros podem ser obtidas por *bootstrap*. Escolhas usuais para $F(.)$ são as funções de distribuição da normal e da logística.

De bastante apelo na modelagem de resultados fracionários é o modelo *two-part*, que postula a distribuição de probabilidades seguinte para $0<\phi_r^*\leq 1$:

$$\phi_r^*(x_r,y_r\mid z_r)=\begin{cases}1 & \text{com probabilidade } F(z_r\beta')\\ H(z_r\delta') & \text{com probabilidade } 1-F(z_r\beta')\end{cases}$$

Portanto, $E\left(\phi_r^*(x_r, y_r \,|z_r)\right) = F(z_r\beta') + (1 - F(z_r\beta'))H(z_r\delta')$.

Esta equação se presta à avaliação de variáveis endógenas, como em Souza et al. (2018), com o uso do GMM ou, sob separabilidade, pode ser estimada por mínimos quadrados não lineares, uma vez que o resíduo tem média zero.

4. RESULTADOS ESTATÍSTICOS

A medida de eficiência DEA varia significativamente por regiões e não suporta a hipótese de retornos constantes à escala – CRS (p-valor = 0,01), conforme teste estatístico via *bootstrap* de Simar e Wilson (2002), disponível no *software* R (Simm e Besstremyannaya, 2016). A aproximação binomial da estatística 'número de ocorrências nas quais CRS<VRS' tem p-valor < 0,0001. A mediana da distribuição é 0,529, coincidente com a média. As medianas regionais são 0,845, 0,686, 0,656, 0,432 e 0,244 para as regiões Centro-oeste, Sudeste, Sul, Norte e Nordeste, respectivamente. Não se observa simetria nas distribuições regionais.

A Tabela 1 mostra os resultados obtidos com o Algoritmo 1 de Simar e Wilson (2007). Foram realizadas 2.000 repetições *bootstrap* e eliminadas 23 observações com eficiência unitária e 115 com valores potencialmente atípicos para a medida de eficiência ($\phi_r^* > 25$). A resposta modelada é ϕ^* . Sem o último filtro os Algoritmos 1 e 2, nesse caso, não convergem.

Tabela 1: Inferência estatística com Algoritmo 1.

	Coeficiente	Desvio padrão *bootstrap*	z	P>|z|	Intervalo de confiança a 95%	
Crédito	-26,571	1,607	-16,53	0,000	-29,838	-23,428
Assist. técnica	-10,428	1,167	-8,93	0,000	-12,797	-8,205
Social	-17,448	2,203	-7,92	0,000	-21,623	-13,061
Demográfico	-33,617	2,819	-11,92	0,000	-39,312	-28,136
Ambiental	5,404	2,596	2,08	0,037	0,208	10,411
Gini	-57,707	2,675	-21,57	0,000	-62,881	-52,486
Norte	12,378	3,935	3,15	0,002	5,221	20,816
Nordeste	23,957	3,981	6,02	0,000	16,804	32,568
Sudeste	20,639	3,920	5,26	0,000	13,506	29,137
Sul	16,432	3,916	4,2	0,000	9,268	24,453
Constante	47,210	4,360	10,83	0,000	38,579	55,362
Sigma	4,926	0,139	35,53	0,000	4,641	5,185

A Tabela 2 apresenta os resultados do Algoritmo 2 de Simar e Wilson (2007). Foi modelada a resposta ϕ^* . Foram realizadas 2.000 repetições *bootstrap*. Além das observações eficientes, como no Algoritmo 1 foram eliminadas as 115 observações potencialmente atípicas para a análise em apreço.

Tabela 2: Inferência estatística com Algoritmo 2.

	Coeficiente	Desvio padrão	z	P>|z|	Intervalo de confiança a 95%

		bootstrap				
Crédito	-36,739	2,280	-16,11	0,000	-41,295	-32,404
Assist. técnica	-13,974	1,576	-8,87	0,000	-17,263	-11,062
Social	-21,024	2,896	-7,26	0,000	-26,788	-15,221
Demográfico	-46,530	3,835	-12,13	0,000	-53,939	-38,822
Ambiental	5,588	3,255	1,72	0,086	-1,015	12,237
Gini	-78,246	3,787	-20,66	0,000	-85,492	-70,657
Norte	22,391	5,917	3,78	0,000	11,808	35,366
Nordeste	38,437	6,067	6,34	0,000	27,789	51,607
Sudeste	32,409	5,974	5,43	0,000	21,864	45,369
Sul	24,823	5,945	4,18	0,000	14,464	37,462
Constante	56,337	6,269	8,99	0,000	42,789	67,423
Sigma	6,047	0,175	34,55	0,000	5,684	6,373

Os resultados estatísticos são coincidentes com as duas abordagens no que tange aos sinais e à significância, exceção feita à variável ambiental que é não significativa no Algoritmo 2. Os resultados foram obtidos do procedimento 'simarwilson' disponível no *software* Stata (Badunenko e Tauchmannn, 2018). A modelagem com $1/\phi^*$, como sugerido acima na exposição, elimina a necessidade de desconsiderar observações atípicas adicionalmente às observações eficientes. O procedimento 'simarwilson' do *software* Stata não funciona com essa modelagem na orientação a produto para o Algoritmo 2. A Tabela 3 apresenta os resultados do Algoritmo 1 para $1/\phi^*$ e 2.000 repetições *bootstrap*.

Tabela 3: Inferência estatística com Algoritmo 1 para $1/\phi^*$.

	Coeficiente	Desvio padrão *bootstrap*	z	P>|z|	Intervalo de confiança a 95%	
Crédito	0,401	0,014	28,27	0,000	0,374	0,428
Assist. técnica	0,295	0,016	18,73	0,000	0,265	0,327
Social	0,337	0,026	12,88	0,000	0,287	0,389
Demográfico	0,324	0,031	10,38	0,000	0,261	0,386
Ambiental	-0,150	0,040	-3,71	0,000	-0,229	-0,070
Gini	1,405	0,034	40,84	0,000	1,337	1,472
Norte	-0,012	0,018	-0,69	0,491	-0,045	0,023
Nordeste	-0,091	0,017	-5,47	0,000	-0,124	-0,059
Sudeste	-0,059	0,016	-3,73	0,000	-0,089	-0,029
Sul	-0,108	0,016	-6,54	0,000	-0,141	-0,076
Constante	-1,121	0,039	-28,74	0,000	-1,201	-1,043
Sigma	0,162	0,002	72,15	0,000	0,157	0,166

Exceto pela significância da *dummy* representativa da diferença entre as regiões Norte e Centro-oeste, os resultados são coincidentes com os da Tabela 1 do ponto de vista das implicações estatísticas.

Para a regressão fracionária optou-se pelo modelo probit, que apresenta verossimilhança maior que a do modelo logístico para a aplicação em apreço. O modelo foi

estimado com uso do *software* Stata (procedimento 'fracreg'). A Tabela 4 mostra os resultados. Os desvios foram determinados por meio de 2.000 repetições *bootstrap*.

Faz-se mister observar a similaridade dos resultados da regressão fracionária com os do Algoritmo 1 para $1/\phi^*$. Têm-se a mesma ordem de grandeza relativa dos parâmetros e a mesma significância estatística.

Para o modelo *two-part* modelaram-se dois construtos lineares distintos para respostas eficientes e não eficientes. Ambos os construtos são compostos com as mesmas variáveis contextuais utilizadas nos modelos anteriores. O número grande de observações facilita esta modelagem que conta, assim, com 22 parâmetros. O modelo foi ajustado por mínimos quadrados não lineares e o quadrado da correlação entre valores preditos e observados é de 75,6%. A Tabela 5 mostra os resultados desta abordagem.

Tabela 4: Inferência estatística com modelo de regressão fracionária probit.

	Coeficiente	Desvio padrão *bootstrap*	z	P>|z|	Intervalo de confiança a 95%	
Crédito	0,903	0,038	23,99	0,000	0,829	0,977
Assist. técnica	0,659	0,041	15,95	0,000	0,578	0,740
Social	0,729	0,060	12,18	0,000	0,611	0,846
Demográfico	0,786	0,071	11,07	0,000	0,647	0,925
Ambiental	-0,422	0,106	-3,96	0,000	-0,631	-0,213
Gini	3,205	0,082	38,96	0,000	3,044	3,366
Norte	-0,038	0,043	-0,87	0,386	-0,123	0,048
Nordeste	-0,241	0,041	-5,93	0,000	-0,320	-0,161
Sudeste	-0,144	0,034	-4,22	0,000	-0,211	-0,077
Sul	-0,234	0,037	-6,4	0,000	-0,305	-0,162
Constante	-3,622	0,096	-37,53	0,000	-3,811	-3,433

Tabela 5: Resultados inferenciais do modelo *two-part*.

	Coeficiente	Desvio padrão	Valor t	Pr > |t|
Respostas eficientes				
Crédito	1,799	0,094	19,14	<0,0001
Assist. técnica	0,570	0,057	9,93	<0,0001
Social	1,018	0,089	11,42	<0,0001
Demográfico	0,512	0,108	4,76	<0,0001
Ambiental	-0,121	0,151	-0,80	0,423
Gini	2,215	0,163	13,62	<0,0001
Norte	-0,401	0,083	-4,81	<0,0001
Nordeste	-0,131	0,066	-2,01	0,045
Sudeste	-0,145	0,059	-2,45	0,014
Sul	-0,335	0,061	-5,49	<0,0001
Constante	-3,664	0,134	-27,39	<0,0001
Respostas ineficientes				
Crédito	-2,676	0,378	-7,09	<0,0001
Assist. técnica	0,495	0,187	2,65	0,008
Social	-1,110	0,377	-2,95	0,003
Demográfico	2,868	0,410	7,00	<0,0001
Ambiental	-1,745	0,458	-3,81	0,000
Gini	8,459	0,459	18,44	<0,0001
Norte	1,154	0,517	2,23	0,026
Nordeste	-0,226	0,507	-0,45	0,656
Sudeste	0,440	0,507	0,87	0,385
Sul	0,404	0,521	0,78	0,438
Constante	-8,175	0,672	-12,16	<0,0001

O aspecto interessante do modelo *two-part* é a modelagem da probabilidade de um município tornar-se eficiente. As medianas regionais são 0,793, 0,639, 0,608, 0,214 e 0,148

para as regiões Centro-oeste, Sul, Sudeste, Norte e Nordeste, respectivamente. O valor mediano da população é 0,403. A Figura 1 ilustra a distribuição regional.

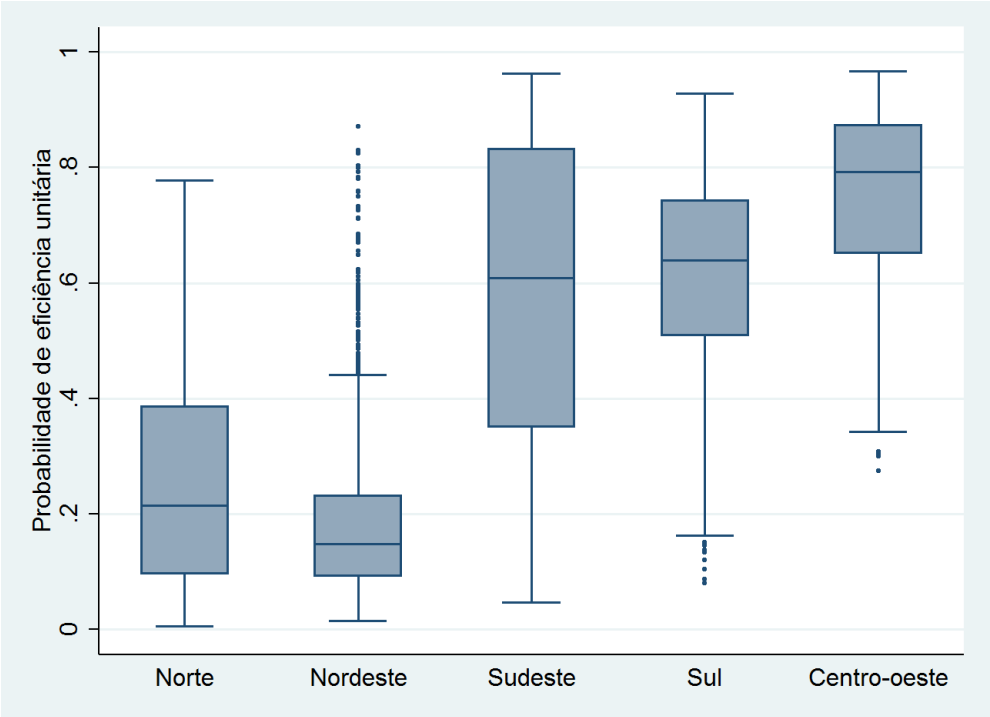

Figura 1: Probabilidade de eficiência unitária, por região.

No construto $z\beta'$ os sinais dos parâmetros coincidem com os do modelo de regressão fracionária. Significância distinta é obtida para o indicador ambiental e a diferença entre as regiões Norte e Centro-oeste, que passa a ser significativa. O construto $z\delta'$, que modela os municípios ineficientes, tem comportamento diferente. Esses municípios não são efetivos na utilização do crédito rural. As condições sociais e o controle ambiental não são favoráveis à produção agropecuária desses municípios.

De um modo geral, a assistência técnica é positiva para a produção e a significância do índice de concentração da renda (Gini) em todos os modelos é indicativo da maior eficiência dos municípios dominantes na produção, que têm a mesma altamente concentrada em poucos estabelecimentos.

A utilização do modelo *two-part*, como aqui especificado para o teste de adequabilidade na presença de variáveis endógenas com o uso do GMM, vai exigir a presença de pelo menos 22 instrumentos. Esses não estão disponíveis na presente aplicação. Veja Souza et al. (2018) para um exemplo de uso do GMM em contexto similar.

No ajuste de mínimos quadrados não lineares, o teste da razão de verossimilhança (Gallant, 1987) rejeita a hipótese $\beta = \delta$ (p-valor < 0,0001).

5. CONSIDERAÇÕES FINAIS

Neste artigo fez-se uma revisão dos principais métodos utilizados para avaliação inferencial de efeitos contextuais em modelos de regressão em dois estágios, com respostas DEA calculadas no primeiro estágio. Consideraram-se os dois algoritmos *bootstrap* de Simar e Wilson e os modelos de regressão fracionária, incluindo a extensão *two-part*, a qual modela diferentemente as respostas DEA eficientes e não eficientes.

De um modo geral, os resultados diferem. Nos algoritmos de Simar e Wilson faz diferença se a modelagem é feita com a medida no intervalo (0,1] ou no intervalo [1,+∞), particularmente com a medida orientada a *output*, onde as duas considerações são viáveis. Entende-se como uma desvantagem desses métodos a eliminação de observações eficientes da análise. Parece ser intuitivo coletar dependências de covariáveis no processo produtivo dos *benchmarks* da amostra. Outro problema encontrado foi a falta de flexibilidade do Algoritmo 2 na modelagem [1,+∞), no procedimento 'simarwilson' disponível no *software* Stata para observações com ineficiência elevada (>25 na aplicação aqui estudada).

Se o número de observações é suficientemente grande, o modelo *two-part* tem bastante apelo. Permite modelar diferentemente observações eficientes e não eficientes e, consequentemente, o teste estatístico da igualdade dos efeitos postulados supostamente atuando nas duas representações por meio de construtos lineares com coeficientes distintos. Além disso, o modelo *two-part* pode ser utilizado na presença de variáveis endógenas com o uso do GMM. A especificação *two-part* também é compatível com a modelagem beta inflacionada (Ospina e Ferrary, 2012).

A aplicação utilizada na comparação das diferentes técnicas para uso de DEA em dois estágios diz respeito à avaliação do efeito de variáveis contextuais potencialmente associadas a imperfeições de mercado na eficiência técnica da produção agropecuária municipal, calculada sob a hipótese de retornos variáveis à escala. Por imperfeições de mercado entende-se uma assimetria nas relações econômicas enfrentadas pelos estabelecimentos rurais. Dificuldades no acesso ao crédito, educação, infraestrutura e informação tornam os preços de insumos desfavoráveis para a pequena produção face às condições de preço que podem ser obtidas na venda de sua produção.

Explorando os resultados do modelo *two-part* conclui-se que o indicador ambiental, embora com efeito negativo, não afeta significativamente os municípios eficientes. As demais covariáveis tem associação positiva e significante. No geral, os sinais dos parâmetros coincidem com os do modelo de regressão fracionária e do Algoritmo 1 quando a eficiência é calculada no intervalo (0,1]. O ajuste dos municípios ineficientes indica que esses municípios não são efetivos na utilização do crédito rural e a importância do indicador de Gini é o reflexo da concentração da renda em poucos estabelecimentos, em geral mais eficientes. As condições sociais e o controle ambiental também não são favoráveis à produção agropecuária desses municípios, refletindo o problema da imperfeição de mercado. Portanto, o caminho das políticas públicas para eliminar as desigualdades nos campos passa pela remoção das imperfeições de mercado. Observa-se que a assistência técnica é positiva para a produção. A região Centro-oeste possui maior probabilidade de municípios, na mediana, tornaram-se eficientes, seguindo-se as regiões Sul, Sudeste, Norte e Nordeste, nesta ordem. Este resultado pode indicar que as políticas públicas podem ser mais efetivas se forem implementadas atendendo as especificidades regionais.

6. REFERÊNCIAS

[1] BĂDIN, L.; DARAIO, C.; SIMAR, L. How to Measure the Impact of Environmental Factors in a Nonparametric Production. European Journal of Operational Research, v. 223, p. 818–833, 2012.

[2] BADUNENKO, O.; TAUCHMANN, H. Simar and Wilson two-stage efficiency analysis for Stata. FAU Discussion Papers in Economics, n. 08/2018. Friedrich-Alexander-Universität Erlangen-Nümberg, Institute for Economics, Erlangen. 37p. 2018.

[3] BANKER, R.D.; CHARNES, A.; COOPER, W.W. Some models for estimating technical scale inefficiencies in Data Envelopment Analysis. Management Science, v. 30, n. 9, p. 1078–1092, 1984.

[4] BANKER, R.D.; NATARAJAN, R. Evaluating contextual variables affecting productivity using data envelopment analysis. Operations Research, v. 56, p. 48–58, 2008.

[5] BANKER, R.D.; NATARAJAN, R. Statistical tests based on DEA efficiency scores. In: W.W. Cooper, L.M. Seiford, & J. Zhu (Eds.), Handbook on Data Envelopment Analysis. Boston: Kluwer International Series, 2011, p. 273–296.

[6] DARAIO, C.; SIMAR, L. Advanced Robust and Nonparametric Methods in Efficiency Analysis. New York: Springer, 2007. 248p.

[7] GALLANT, A.R. Nonlinear Statistical Models. New York: Wiley, 1987. 610p.

[8] GOMES, E.G.; SOUZA, G.S.; VIVALDI, L.J. Two-stage inference in experimental design using DEA: an application to intercropping and evidence from randomization theory. Pesquisa Operacional, v. 28, p. 339–354, 2008.

[9] OECD/FAO. OCDE-FAO Perspectivas Agrícolas 2015, OECD Publishing, Paris. 2015. Disponível em: http://www.fao.org/3/a-i4738s.pdf. Acesso em: 01 Fev. 2019.

[10] OSPINA, R.; FERRARI, S.L.P. A general class of zero-or-one inflated beta regression models. Computational Statistics and Data Analysis, v. 56, n. 6, p. 1609–1623, 2012.

[11] PAPKE, L.E.; WOOLDRIDGE, J.M. Econometric methods for fractional response variables with an application to 401(k) plan participation rates. Journal of Applied Economics, v. 11, n. 6, p. 619–632, 1996.

[12] RAMALHO, E.A.; RAMALHO, J.J.S.; HENRIQUES, P.D. Fractional regression models for second stage DEA efficiency analyses. Journal of Productivity Analysis, v. 34, p. 239–255, 2010.

[13] SIMAR, L.; WILSON, P.W. Estimation and inference in two-stage, semi-parametric models of production processes. Journal of Econometrics, v. 136, n. 1, p. 31–64, 2007.

[14] SIMAR, L.; WILSON, P.W. Non-parametric tests of returns to scale. European Journal of Operational Research, v. 139, n. 1, p. 115–132, 2002.

[15] SIMAR, L.; WILSON, P.W. Sensitivity analysis of efficiency scores: how to *bootstrap* in nonparametric frontier models. Management Science, v. 44, n. 1, p. 49–61, 1998.

[16] SIMAR, L.; WILSON, P.W. Two-stage DEA: caveat emptor. Journal of Productivity Analysis, v. 36, n. 2, p. 205–218, 2011.

[17] SIMM, J.; BESSTREMYANNAYA, G. Robust Data Envelopment Analysis (DEA) for R. Repository CRAN, 2016. Disponível em: https://github.com/jaak-s/rDEA. Acesso em: 01 Fev. 2019.

[18] SOUZA, G.S.; ALVES, E.R.A.; GOMES, E.G.; MAGALHÃES, E.; ROCHA, D.P. Um modelo de produção para a agricultura brasileira e a importância da pesquisa da Embrapa. In: Alves, E.R.A.; Souza, G.S.; Gomes, E.G. (Org.). Contribuição da Embrapa para o desenvolvimento da agricultura no Brasil. 1ed.Brasília: Embrapa Informação Tecnológica, 2013, v. 1, p. 49–86.

[19] SOUZA, G.S.; GOMES, E.G. Management of agricultural research centers in Brazil: A DEA application using a dynamic GMM approach. European Journal of Operational Research, v. 240, p. 819–824, 2015.

[20] SOUZA, G.S.; GOMES, E.G.; ALVES, E.R.A. Determinantes da dispersão da renda no meio rural brasileiro. Blucher Marine Engineering Proceedings, v. 2, p. 173–184, 2016.

[21] SOUZA, G.S.; GOMES, E.G.; ALVES, E.R.A. Imperfeições de mercado e concentração de renda na produção agrícola. Revista de Política Agrícola, v. 27, p. 31–38, 2018.