Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Dissertação de Mestrado

Endogeneidade em fronteiras estocásticas de produção com modelos de um e dois estágios: uma aplicação com dados do Censo Agropecuário

por

Kessys Lorrânya Peralta de Oliveira

Brasília, Dezembro de 2019

Endogeneidade em fronteiras estocásticas de produção com modelos de um e dois estágios: uma aplicação com dados do Censo Agropecuário

por

Kessys Lorrânya Peralta de Oliveira

Dissertação apresentada ao Departamento de Estatística da Universidade de Brasília, como requisito parcial para obtenção do título de Mestre em Estatística.

Orientador: Prof. Dr. Bernardo Borba de Andrade

Coorientador: Prof. Dr. Geraldo da Silva e Souza

Brasília, Dezembro de 2019

Dissertação submetida ao Programa de Pós-Graduação em Estatística do Departamento de Estatística da Universidade de Brasília como parte dos requisitos para a obtenção do grau de Mestre em Estatística.

Texto aprovado por:

Prof. Bernardo Borda de Andrade
Orientador, EST/UnB

Prof. Geraldo da Silva e Souza
Coorientador, Embrapa/UnB

Dr.ª Eliane Gonçalves Gomes
Embrapa

Prof. André Luiz Fernandes Cançado
EST/UnB

The greater our knowledge increases the more our ignorance unfolds.

(John F. Kennedy)

Dedico este trabalho aos meus pais e aos meus irmãos, com quem compartilhei momentos de alegria, tristeza e ansiedade. Ao meu namorado, por toda compreensão, carinho e amor. Aos meus amigos, Bruno e Jean, que estiveram ao meu lado durante esta longa caminhada. Aos demais amigos e colegas, que me incentivaram e ofereceram apoio nos momentos críticos. A todos os membros do Departamento de Estatística da UnB, e as pessoas com quem convivi nesses espaços ao longo desses dois anos.

Meus sinceros agradecimentos aos professores do PPGEST/UnB, em especial, aos professores Bernardo Borba e Geraldo da Silva por toda a paciência e dedicação despendida para que esse trabalho se concretizasse e aos demais professores do Departamento de Estatística, por todo conhecimento adquirido e pelo exemplo de pessoas que são.

Agradeço aos meus colegas de curso Bruno, Jean, Leonardo, Regina, José Paulo, Luana, Rubens e Lais que sempre estiveram dispostos a ajudar. Em especial, ao Bruno, que esteve comigo desde o início, um ajudando o outro no aprendizado das disciplinas e na elaboração da dissertação, principalmente na implementação dos códigos em *R*.

Meus agradecimentos a toda minha família e ao meu namorado que sempre me apoiaram e ajudaram em momentos de dificuldade e desânimo. Em especial, ao meu irmão Gustavo, o qual me motivou e me serviu de inspiração por sua coragem e dedicação em tudo que faz.

Agradeço a todos os servidores do Departamento de Estatística, por serem pessoas tão gentis e prestativas.

Agradeço ao Instituto Federal de Rondônia pelo afastamento integral concedido e aos meus colegas de trabalho do IFRO.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Resumo

Modelos de fronteira estocástica de produção são amplamente utilizados em microeconometria e, nas últimas décadas, têm se mostrado bastantes versáteis quanto a sua aplicação. Contudo, existem poucos estudos que tratam da endogeneidade em modelos de fronteira estocástica de produção. Nesta perspectiva, o presente trabalho apresenta dois modelos de fronteira estocástica de produção com variáveis endógenas baseados nas principais distribuições para o termo de ineficiência técnica: as distribuições seminormal, exponencial e normal truncada. A metodologia apresentada aqui é baseada na estimação por máxima verossimilhança de um e dois estágios e é implementada na linguagem *R*. Ao final, uma aplicação aos dados municipais do censo agropecuário brasileiro de 2006 é feita, incluindo vários aspectos de estimação e inferência.

Palavras-chave: Endogeneidade; Fronteira estocástica de produção; Ineficiência técnica; Linguagem *R*; Método de máxima verossimilhança.

Abstract

Stochastic production frontier models are widely used in microeconometrics and, in the last decades, have been proved versatile in their range of applications. There are few studies dealing with 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: the half normal, exponential and truncated normal distributions. The methodology presented here is based on one and two-stage maximum likelihood estimation and it is implemented in *R* language. We also present an application with municipal data from the 2006 Brazilian agricultural census including several aspects of estimation and inference.

Keywords: Endogeneity; Maximum likelihood method; *R* Language; Stochastic production frontier; Technical inefficiency.

Sumário

Lista de Tabelas

Lista de Figuras

Abreviações e Siglas

BFGS	Método Broyden-Fletcher-Goldfarb-Shanno
BHHH	Algoritmo Berndt-Hall-Hall-Hausman
CV	Coeficiente de variação
EP	Erro padrão
Eq.	Equação
IBGE	Instituto Brasileiro de Geografia e Estatística
iid	Independente e identicamente distribuída
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
LI	Limite inferior do intervalo de confiança
LS	Limite superior do intervalo de confiança
MVIC	Máxima verossimilhança de informação completa
MVIL	Máxima verossimilhança de informação limitada
P-valor	Nível descritivo de um teste de hipóteses
QMO	Quadrados mínimos ordinários
REQM	Raiz do erro quadrático médio
R	Linguagem *R*
TRV	Teste da razão de verossimilhança

Lista de Símbolos e Notações

$\mathrm{Cor}(\cdot)$	correlação
$\mathrm{Cov}(\cdot)$	covariância
∂	diferencial parcial
$\mathcal{N}$	distribuição normal
χ^2	estatística qui-quadrado
F	estatística *F-Snedecor*
t	estatística *t-Student*
z	estatística z
$\mathcal{E}$	eficiência técnica
f_u	função de densidade de u
$\phi(\cdot)$	função de densidade normal padrão
$\Phi(\cdot)$	função de distribuição acumulada normal padrão
H_0	hipótese nula
H_1	hipótese alternativa
$\perp$	independência
i, j	índices
$\ln L$	log-verossimilhança
n	tamanho amostral
u	termo de erro unilateral / ineficiência técnica
v	termo de erro bilateral / idiossincrático
$^\top$	transposto
$\mathrm{E}(\cdot)$	valor esperado
$\mathrm{Var}(\cdot)$	variância

Capítulo 1

Introdução

Os modelos de fronteira estocástica são amplamente utilizados em microeconometria. Um modelo de fronteira estocástica é um modelo de efeitos aleatórios projetado para fornecer estimativas da eficiência técnica de unidades tomadoras de decisão, ou ainda, de produtores, por meio de funções de produção ou de custo. A literatura possui muitos exemplos empíricos de vários campos, tais como agricultura, bancos e saúde.

A análise de fronteira estocástica pressupõe que cada produtor produz potencialmente menos do que poderia devido a um grau de ineficiência. Esta ineficiência é incorporada ao modelo por meio de uma variável aleatória $u_i \geq 0$, que representa a ineficiência técnica do i-ésimo produtor. Supõe-se também que a produção do i-ésimo produtor está sujeita à ruídos aleatórios, $v_i \in \mathbb{R}$, que refletem a característica estocástica da fronteira, definido por erro aleatório que afeta o processo produtivo. Assume-se que u_i e v_i são independentes entre si e dos regressores, conhecidos como insumos, no modelo de fronteira estocástica de produção.

Os parâmetros dos modelos de fronteira estocástica são tipicamente estimados por método de máxima verossimilhança ou por quadrados mínimos ordinários corrigidos, em que a consistência de cada um destes estimadores depende da exogeneidade das variáveis explicativas.

Em trabalhos empíricos é comum a suposição de que u_i tem distribuição seminormal[1]. No entanto, outras suposições distribucionais sobre o componente de erro unilateral, u_i, foram

[1] Conforme a Subseção 2.2.1

propostas, como a distribuição exponencial[2] e a normal truncada[3], e elas também foram empregadas, embora com menos frequência, em trabalhos empíricos.

Há implementações disponíveis em alguns programas para a análise de fronteira estocástica. No entanto, muitas das implementações existentes lidam apenas com a suposição de exogeneidade, não tratando a endogeneidade que pode existir quando uma ou mais das covariáveis da fronteira ou da ineficiência forem correlacionadas com o termo de erro bilateral, v_i.

Além disso, muitos programas utilizam procedimentos numéricos para estimação dos parâmetros. Uma forma de tornar a velocidade de convergência do procedimento de otimização iterativa consideravelmente melhor, é incluir o vetor de gradientes analíticos nesse processo de estimação. Na literatura de análise de fronteira estocástica, na presença de covariáveis endógenas, ainda não há disponível trabalhos que ilustram as expressões desses gradientes.

1.1 Justificativa

A maior parte da literatura que lida com a presença de endogeneidade em modelos de fronteira estocástica, consideram a suposição distributiva seminormal para o termo de ineficiência. Portanto, devido à baixa quantidade de trabalhos empíricos sobre outras suposições distribucionais para o termo de ineficiência, principalmente ao lidar com a endogeneidade, este trabalho aborda, além da distribuição seminormal, outras distribuições para u_i, com o uso de dados de corte transversal. Especificamente, apresentam-se técnicas de modelagem da ineficiência técnica em modelos de fronteira estocástica de produção na presença de endogeneidade, com o uso de diferentes variações de estimação por máxima verossimilhança, incluindo as expressões dos gradientes analíticos desses modelos. Com a implementação destas técnicas em linguagem *R* que não possui pacote ainda disponível para lidar com a endogeneidade na análise de fronteira estocástica de forma mais ampla.

[2]Conforme a Subseção 2.2.2
[3]Conforme a Subseção 2.2.3

1.2 Objetivo

Este trabalho tem o objetivo de implementar a predição do nível de eficiência técnica de produtores, em fronteiras estocásticas de produção na presença de variáveis endógenas, via máxima verossimilhança em um e dois estágios, com o uso de diferentes especificações do termo de ineficiência, bem como, expressar os gradientes analíticos desses modelos. Para tanto, irá percorrer-se as seguintes etapas:

- Ilustrar as especificações seminormal, exponencial e normal truncada do termo de ineficiência, nos modelos de fronteira estocástica; (Cap. 2)
- Usar modelos para lidar com a endogeneidade por método de máxima verossimilhança em um estágio baseado em Karakaplan e Kutlu (2015); (Cap. 3)
- Usar modelos para lidar com a endogeneidade por método de máxima verossimilhança em dois estágios com a correção da variância de Murphy e Topel (1985); (Cap. 4)
- Predizer a eficiência técnica de produtores em modelos de fronteira estocástica de produção na presença de endogeneidade; (Cap. 5)
- Mostrar alguns testes de endogeneidade; (Cap. 6)
- Apresentar uma aplicação a dados reais; (Cap. 7)
- Disponibilizar a rotina implementada em linguagem *R*.

1.3 Dados

Para fins da aplicação são usados dados de corte transversal provenientes do censo agropecuário do IBGE de 2006, agregados em nível municipal, além de dados do censo demográfico brasileiro de 2010 e das bases de dados do INEP, referentes à educação em 2009, e dos dados do Ministério da Saúde de 2011. Dados esses válidos para 4965 municípios, que representam quase 90% do total de municípios brasileiros.

1.4 Esboço do trabalho

Este trabalho inicia-se com uma breve revisão sobre os modelos de fronteira estocástica. O Capítulo 2 descreve a análise de fronteira estocástica de produção em modelos com regressores exógenos ou endógenos. O Capítulo 3 trata de um método de um estágio para estimação de fronteiras estocásticas de produção na presença da endogeneidade, com diferentes especificações do termo de ineficiência. O Capítulo 4 aborda um método de dois estágios para estimação em modelos de fronteira estocástica com variáveis endógenas. O Capítulo 5 trata da predição de eficiência técnica da unidade de produção. O Capítulo 6 discute testes de hipóteses e ajuste de modelos. O Capítulo 7 apresenta os resultados de uma aplicação das técnicas a dados reais. Por fim, o Capítulo 8 trata das conclusões obtidas, bem como possíveis trabalhos futuros.

Capítulo 2

Fronteira Estocástica de Produção

O objetivo deste capítulo é apresentar os principais modelos propostos para o termo de ineficiência na análise de fronteira estocástica de produção e discutir sobre a presença de endogeneidade.

2.1 Considerações iniciais

Modelos de fronteira estocástica de produção foram simultaneamente introduzidos por Aigner, Lovell e Schmidt (1977) e Meeusen e Den Broeck (1977). Com dados de corte transversal sobre as quantidades de k insumos utilizados para produzir um único produto para cada um dos n produtores, um modelo de fronteira estocástica de produção pode ser escrito como

$$y_i = f(x_i;\beta)\exp(v_i)\mathcal{E}_i = f(x_i;\beta)\exp(v_i - u_i), \quad e_i = v_i - u_i, \quad i = 1, 2, \ldots, n, \qquad (2.1)$$

em que x_i é um vetor de k insumos utilizados pelo i-ésimo produtor, β é um vetor $k \times 1$ de parâmetros tecnológicos a serem estimados e $[f(x_i;\beta)\exp(v_i)]$ é a fronteira de produção estocástica que consiste em duas partes: uma parte determinística, $f(x_i;\beta)$, comum a todos os produtores e uma parte específica do produtor, $\exp(v_i)$, que captura o efeito de choques aleatórios específicos de cada produtor.

Este modelo é preferido ao modelo que não considera choques aleatórios porque o modelo anterior corre o risco de atribuir incorretamente variação ambiental não modelada à variação na

eficiência técnica (Kumbhakar e Lovell, 2003).

Desde que a componente $\mathcal{E}_i = \exp(-u_i)$ é a eficiência técnica orientada para a produção do i-ésimo produtor, tem-se

$$\mathcal{E}_i = \frac{y_i}{f(x_i;\beta)\exp(v_i)}, \quad 0 < \mathcal{E}_i < 1, \tag{2.2}$$

que define a eficiência técnica como a razão entre a produção observada e à produção máxima viável, isto é, $\mathcal{E}_i$ fornece uma medida do déficit da produção observada de cada produtor em relação a produção máxima viável em um ambiente caracterizado por $\exp(v_i)$ (Kumbhakar e Lovell, 2003).

As estimativas da eficiência específica de cada produtor dependem da decomposição de e_i e tipicamente são derivadas da esperança condicional de $\exp(-u_i)$ dado e_i, que variam conforme as funções densidade de probabilidade de ambos v_i e u_i.

2.2 Modelos de fronteira estocástica de produção

Considere o modelo de fronteira estocástica de produção da Eq. (2.1) na forma log-linear de Cobb-Douglas,

$$\ln y_i = \beta_0 + \sum_{j=1}^{k} \beta_j \ln x_{ji} + v_i - u_i. \tag{2.3}$$

Nesta especificação, a $\ln f(x_i;\beta)$ produz um modelo linear em β e o modelo da Eq. (2.3) é, portanto, um modelo de regressão com o usual ruído gaussiano, $v \sim \mathcal{N}(0,\sigma_v^2)$, e um efeito aleatório, $u \sim f_u$, representando a ineficiência técnica da unidade.

Os dados utilizados para se estimar uma fronteira de produção consistem em observações sobre a quantidade de insumos empregados, $\boldsymbol{x_i}$, e a produção obtida por cada produtor, y_i. Nenhuma informação de preço é usada, e nenhum objetivo comportamental é imposto aos produtores. As técnicas de estimativa empregadas dependem, em parte, da riqueza da quantidade de dados disponíveis. O interesse principal é a estimativa da eficiência técnica, presumindo que os produtores produzem apenas um único produto, seja porque realmente produzem apenas um

único produto ou porque é possível agregar seus múltiplos produtos em um índice de produto único (Kumbhakar e Lovell, 2003).

O termo de erro composto $e_i = v_i - u_i$ é assimétrico, desde que $u_i \geq 0$. A suposição de que u_i e v_i são independentes entre si e dos regressores é problemática, uma vez que, se os produtores souberem algo sobre sua eficiência técnica, isso pode influenciar suas escolhas de insumos. A estimação por quadrados mínimos ordinários (QMO) fornece estimativas consistentes dos parâmetros $\boldsymbol{\beta_j}$, mas não de β_0, dado que $\mathrm{E}(e_i) = -\mathrm{E}(u_i) \leq 0$. Além disso, não fornece estimativas consistentes da eficiência técnica específica de cada produtor (Kumbhakar e Lovell, 2003).

Uma solução para esse problema é corrigir o viés no intercepto usando o estimador de quadrados mínimos ordinários corrigidos (COLS, em inglês) ou o estimador de quadrados mínimos ordinários modificados (MOLS, em inglês). Uma solução discutivelmente melhor é fazer algumas suposições distribucionais relativas aos dois termos de erro e estimar os parâmetros do modelo pelo método de máxima verossimilhança. Uma vez que este possui muitas propriedades estatísticas desejáveis em grandes amostras (isto é, assintóticas), são frequentemente preferidos a outros estimadores tais como o COLS ou o MOLS (Coelli et al., 2005).

Portanto, o método de máxima verossimilhança pode ser usado para estimar os parâmetros do modelo. Neste contexto, o termo de ineficiência é uma variável latente que deve ser integrada ao calcular a verossimilhança. Dependendo da escolha da função de densidade de u_i, o cálculo da verossimilhança exigirá integração numérica (Andrade e Souza, 2017). As especificações mais usuais para o componente de ineficiência técnica das unidades, u_i, são:

Seminormal (Aigner, Lovell e Schmidt, 1977): $u_i \sim$ iid $\mathcal{N}^+(0, \sigma_u^2)$;

Exponencial (Aigner, Lovell e Schmidt, 1977; Meeusen e Den Broeck, 1977): $u_i \sim$ iid $\mathrm{Exp}(\sigma_u)$;

Normal truncada (Stevenson, 1980): $u_i \sim$ iid $\mathcal{N}^+(\mu, \sigma_u^2)$;

Gama (Stevenson, 1980; Greene, 1990): $u_i \sim$ iid $\mathrm{Gama}(\sigma_u, m)$.

2.2.1 Modelo normal/seminormal

Considere o modelo de fronteira estocástica de produção dado na Eq. (2.3), assumindo as seguintes hipóteses:

(i) $v_i \sim \text{iid}\, \mathcal{N}(0, \sigma_v^2)$.

(ii) $u_i \sim \text{iid}\, \mathcal{N}^+(0, \sigma_u^2)$.

(iii) $v_i \perp u_i$ e $(v_i, u_i) \perp \boldsymbol{x}_i$.

Dada a hipótese de independência, a função de densidade conjunta de u e v é o produto de suas funções de densidade individuais e, como $e = v - u$, a função de densidade conjunta para u e e é

$$f(u, e) = \frac{2}{2\pi\sigma_u\sigma_v} \exp\left\{-\frac{u^2}{2\sigma_u^2} - \frac{(e+u)^2}{2\sigma_v^2}\right\}. \tag{2.4}$$

A função de densidade marginal de e é obtida pela integração de $f(u, e)$ em u, que produz

$$f(e) = \frac{2}{\sigma}\phi\left(\frac{e}{\sigma}\right)\Phi\left(-\frac{e\lambda}{\sigma}\right), \tag{2.5}$$

em que $\sigma = (\sigma_u^2 + \sigma_v^2)^{1/2}$ e $\lambda = \sigma_u/\sigma_v$. $\phi(\cdot)$ e $\Phi(\cdot)$ são a função de densidade e a função de distribuição acumulada normal padrão, respectivamente. A reparametrização de σ_u^2 e σ_v^2 para σ e λ é conveniente, uma vez que λ fornece uma indicação da contribuição de variabilidade relativa de u e v para e. Quando $\lambda \to 0$, v domina u na determinação de e, implicando em um modelo de função de produção QMO sem eficiência técnica. Quando $\lambda \to \infty$, u domina v na determinação de e, obtendo um modelo de fronteira de produção determinístico sem ruído.

Logo, a distribuição normal/seminormal possui dois parâmetros, σ_u e σ_v (ou λ e σ), que são estimados juntamente com os parâmetros tecnológicos β. A Figura 2.1 mostra três destas distribuições correspondentes a três combinações de σ_u e σ_v. Note que todas são assimétricas negativas, com moda (e média) negativa, desde que $\sigma_u > 0$.

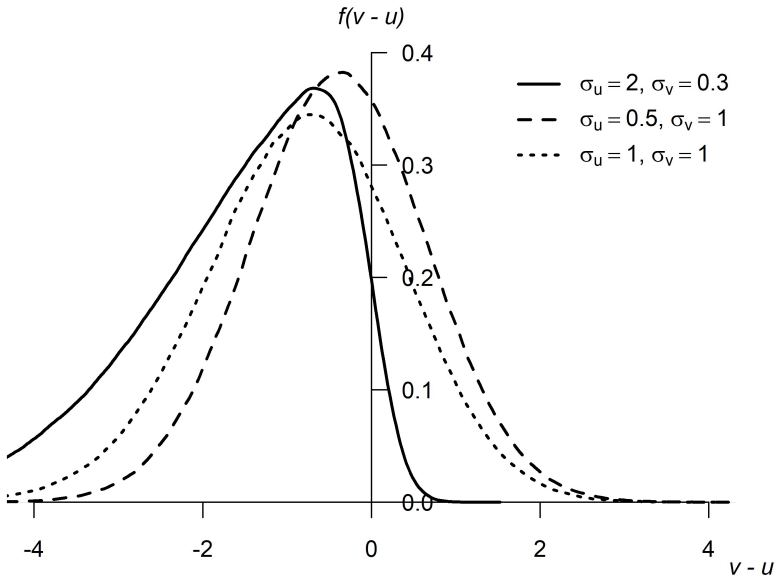

Figura 2.1: Distribuições normal/seminormal

O termo de erro composto, e, é assimetricamente distribuído, com média e variância

$$\begin{aligned} \mathrm{E}(e) &= -\mathrm{E}(u) = -\sigma_u \sqrt{\frac{2}{\pi}}, \\ \mathrm{Var}(e) &= \frac{\pi - 2}{\pi}\sigma_u^2 + \sigma_v^2. \end{aligned} \tag{2.6}$$

Da Eq. (2.5), a função de log-verossimilhança do modelo normal/seminormal é

$$\ln L = \sum_{i=1}^{n} \left\{ \frac{1}{2} \ln \left(\frac{2}{\pi} \right) - \frac{1}{2} \ln \sigma^2 + \ln \Phi \left(-\frac{e_i \lambda}{\sigma} \right) - \frac{e_i^2}{2\sigma^2} \right\}. \tag{2.7}$$

Da Eq. (2.7) obtém-se o gradiente, dado por

$$\begin{aligned} U(\beta) &= \frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n} x_i^\top \left\{ \frac{e_i}{\sigma^2} + \frac{\lambda}{\sigma} A_i \right\}, \\ U(\lambda) &= \frac{\partial \ln L}{\partial \lambda} = \sum_{i=1}^{n} \left\{ -\frac{e_i}{\sigma} A_i \right\}, \\ U(\sigma^2) &= \frac{\partial \ln L}{\partial \sigma^2} = \sum_{i=1}^{n} \left\{ \frac{1}{2\sigma^2} \left[\frac{e_i^2}{\sigma^2} + \frac{e_i \lambda}{\sigma} A_i - 1 \right] \right\}, \end{aligned} \tag{2.8}$$

ou ainda, $U(\beta)$ como acima e

$$\begin{aligned} U(\sigma_u^2) &= \frac{\partial \ln L}{\partial \sigma_u^2} = \sum_{i=1}^{n} \left\{ \frac{1}{2\sigma^2} \left[\frac{e_i^2}{\sigma^2} - \frac{e_i}{\lambda\sigma} A_i - 1 \right] \right\}, \\ U(\sigma_v^2) &= \frac{\partial \ln L}{\partial \sigma_v^2} = \sum_{i=1}^{n} \left\{ \frac{1}{2\sigma^2} \left[\frac{e_i^2}{\sigma^2} + \frac{e_i\lambda}{\sigma}(2+\lambda^2) A_i - 1 \right] \right\}, \end{aligned} \tag{2.9}$$

em que $A_i = \dfrac{\phi(a_i)}{\Phi(a_i)}$ com $a_i = -\dfrac{e_i\lambda}{\sigma}$.

2.2.2 Modelo normal/exponencial

Considere novamente o modelo de fronteira estocástica de produção dado na Eq. (2.3), mas agora assumindo as hipóteses:

(i) $v_i \sim$ iid $\mathcal{N}(0, \sigma_v^2)$.

(ii) $u_i \sim$ iid $\mathrm{Exp}(\sigma_u)$.

(iii) $v_i \perp u_i$ e $(v_i, u_i) \perp \boldsymbol{x}_i$.

A função de densidade conjunta de u e e é

$$f(u, e) = \frac{1}{\sqrt{2\pi}\sigma_u\sigma_v} \exp\left\{ -\frac{u}{\sigma_u} - \frac{(e+u)^2}{2\sigma_v^2} \right\}. \tag{2.10}$$

A função de densidade marginal de e é

$$f(e) = \left(\frac{1}{\sigma_u}\right) \Phi\left(-\frac{e}{\sigma_v} - \frac{\sigma_v}{\sigma_u} \right) \exp\left(\frac{e}{\sigma_u} + \frac{\sigma_v^2}{2\sigma_u^2} \right), \tag{2.11}$$

em que $\Phi(\cdot)$ é a função de distribuição acumulada normal padrão.

A Figura 2.2 ilustra três distribuições normal/exponencial correspondentes a três combinações dos parâmetros σ_u e σ_v. Todas as três são assimétricas negativas. Quando $\sigma_u \to \infty$, a distribuição tende a uma exponencial negativa. No entanto, quando $\sigma_v \to \infty$, a distribuição se parece mais com uma normal.

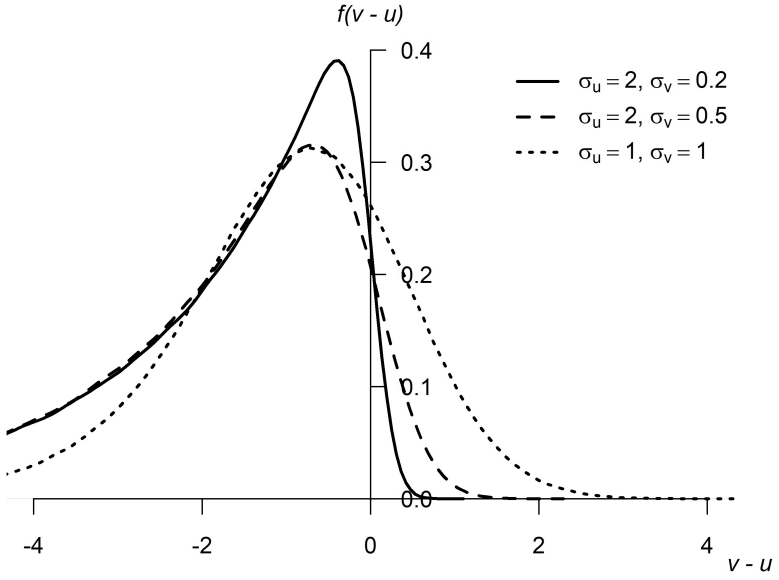

Figura 2.2: Distribuições normal/exponencial

O termo de erro composto, e, é assimetricamente distribuído, com média e variância

$$\begin{aligned} \mathrm{E}(e) &= -\mathrm{E}(u) = -\sigma_u, \\ \mathrm{Var}(e) &= \sigma_u^2 + \sigma_v^2. \end{aligned} \tag{2.12}$$

Da Eq. (2.11), a log-verossimilhança do modelo normal/ exponencial é

$$\ln L = \sum_{i=1}^{n} \left\{ -\frac{1}{2} \ln \sigma_u^2 + \frac{\sigma_v^2}{2\sigma_u^2} + \ln \Phi \left(\frac{-e_i - \frac{\sigma_v^2}{\sigma_u}}{\sigma_v} \right) + \frac{e_i}{\sigma_u} \right\}. \tag{2.13}$$

Da Eq. (2.13) obtém-se o gradiente, dado por

$$\begin{aligned} U(\beta) &= \frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n} x_i^\top \left\{ \frac{1}{\sigma_v} B_i - \frac{1}{\sigma_u} \right\}, \\ U(\sigma_u^2) &= \frac{\partial \ln L}{\partial \sigma_u^2} = \sum_{i=1}^{n} \left\{ \frac{1}{2\sigma_u^2} \left[\frac{\sigma_v}{\sigma_u} B_i - \frac{\sigma_v^2}{\sigma_u^2} - \frac{e_i}{\sigma_u} - 1 \right] \right\}, \\ U(\sigma_v^2) &= \frac{\partial \ln L}{\partial \sigma_v^2} = \sum_{i=1}^{n} \left\{ \frac{1}{2\sigma_u^2} + \frac{1}{2\sigma_v} \left[\frac{e_i}{\sigma_v^2} - \frac{1}{\sigma_u} \right] B_i \right\}, \end{aligned} \tag{2.14}$$

em que $B_i = \dfrac{\phi(b_i)}{\Phi(b_i)}$ com $b_i = \dfrac{-e_i - \sigma_v^2/\sigma_u}{\sigma_v}$.

2.2.3 Modelo normal/normal truncada

Considere novamente o modelo de fronteira estocástica de produção dado na Eq. (2.3), nesta formulação os pressupostos distributivos são:

(i) $v_i \sim$ iid $\mathcal{N}(0, \sigma_v^2)$.

(ii) $u_i \sim$ iid $\mathcal{N}^+(\mu, \sigma_u^2)$.

(iii) $v_i \perp u_i$ e $(v_i, u_i) \perp \boldsymbol{x}_i$.

A distribuição normal truncada assumida para u generaliza a distribuição seminormal de um parâmetro, permitindo que a distribuição normal, que é truncada à esquerda em zero, tenha uma moda diferente de zero.

A função de densidade conjunta de u e e é

$$f(u, e) = \frac{1}{2\pi\sigma_u\sigma_v\Phi(\mu/\sigma_u)} \exp\left\{-\frac{(u-\mu)^2}{2\sigma_u^2} - \frac{(e+u)^2}{2\sigma_v^2}\right\}. \tag{2.15}$$

A função de densidade marginal de e é

$$f(e) = \frac{1}{\sigma}\phi\left(\frac{e+\mu}{\sigma}\right)\Phi\left(\frac{\mu}{\sigma\lambda} - \frac{e\lambda}{\sigma}\right)\left[\Phi\left(\frac{\mu}{\sigma_u}\right)\right]^{-1}, \tag{2.16}$$

em que $\sigma = (\sigma_u^2 + \sigma_v^2)^{1/2}$ e $\lambda = \sigma_u/\sigma_v$. $\phi(\cdot)$ e $\Phi(\cdot)$ são a função de densidade e a função de distribuição acumulada normal padrão, respectivamente. Se $\mu = 0$ a Eq. (2.16) colapsa para a função de densidade marginal do modelo normal/seminormal dado pela Eq. (2.5).

O termo de erro composto, e, é assimetricamente distribuído, com média e variância

$$\begin{aligned} \mathrm{E}(e) &= -\mathrm{E}(u) = -\mu - \sigma_u a, \\ \mathrm{Var}(e) &= \sigma_u^2\left(1 - \frac{\mu}{\sigma_u}a - a^2\right) + \sigma_v^2, \end{aligned} \tag{2.17}$$

sendo $a = \dfrac{\phi\left(\dfrac{\mu}{\sigma_u}\right)}{\Phi\left(\dfrac{\mu}{\sigma_u}\right)}$.

A distribuição normal/normal truncada possui três parâmetros, um parâmetro de locação, $\mu \in \mathbb{R}$, e dois de escala, σ_u e σ_v, (ou μ, λ e σ), que são estimados juntamente com os parâmetros tecnológicos β. A Figura 2.3 mostra três destas distribuições correspondentes a três combinações de μ com $\sigma_u = \sigma_v = 1$. Note que todas são assimétricas negativas, com moda (e média) negativa.

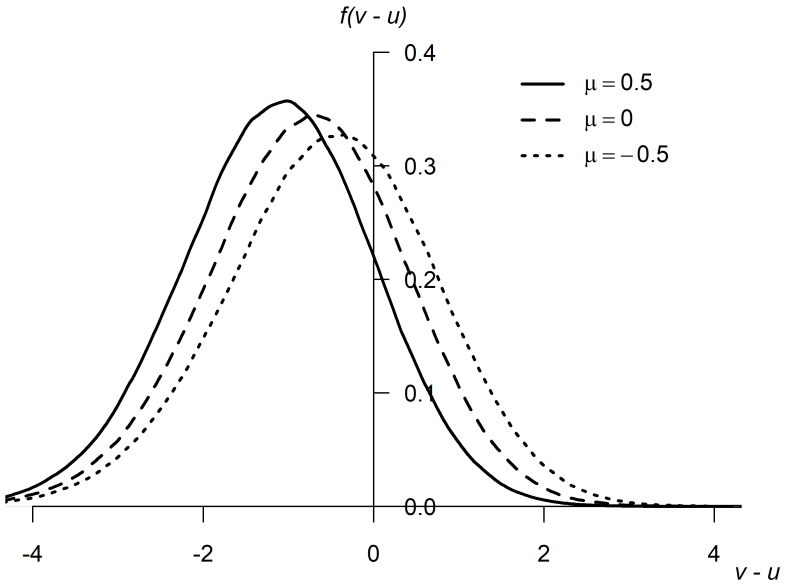

Figura 2.3: Distribuições normal/normal truncada

Da Eq. (2.16), a log-verossimilhança do modelo normal/normal truncada é

$$\ln L = \sum_{i=1}^{n} \left\{ -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\sigma^2 - \ln\Phi\left(\frac{\mu}{\sigma_u}\right) + \ln\Phi\left(\frac{\mu}{\sigma\lambda} - \frac{e_i\lambda}{\sigma}\right) - \frac{1}{2}\left(\frac{e_i+\mu}{\sigma}\right)^2 \right\}, \tag{2.18}$$

sendo $\sigma_u = \sigma(\lambda^{-2}+1)^{-\frac{1}{2}}$.

Da Eq. (2.18) obtém-se o gradiente, dado por

$$\begin{aligned}
U(\beta) &= \frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n} x_i^\top \left\{ \frac{e_i + \mu}{\sigma^2} + \frac{\lambda}{\sigma} D_i \right\}, \\
U(\lambda) &= \frac{\partial \ln L}{\partial \sigma^2} = \sum_{i=1}^{n} \left\{ -\frac{1}{\sigma} \left(\frac{\mu}{\lambda^2} + e_i \right) D_i + \frac{\mu}{\sigma \lambda^3} (\lambda^{-2} + 1)^{-\frac{1}{2}} C \right\}, \\
U(\sigma^2) &= \frac{\partial \ln L}{\partial \lambda} = \sum_{i=1}^{n} \left\{ \frac{1}{2\sigma^2} \left[\frac{(e_i + \mu)^2}{\sigma^2} - \frac{1}{\sigma} \left(\frac{\mu}{\lambda} - e_i \lambda \right) D_i + \frac{\mu \sqrt{\lambda^{-2} + 1}}{\sigma} C - 1 \right] \right\}, \\
U(\mu) &= \frac{\partial \ln L}{\partial \mu} = \sum_{i=1}^{n} \left\{ \frac{1}{\lambda \sigma} D_i - \frac{\sqrt{\lambda^{-2} + 1}}{\sigma} C - \frac{e_i + \mu}{\sigma^2} \right\},
\end{aligned} \tag{2.19}$$

ou ainda, $U(\beta)$ e $U(\mu)$ como acima e

$$\begin{aligned}
U(\sigma_u^2) &= \frac{\partial \ln L}{\partial \sigma_u^2} = \sum_{i=1}^{n} \left\{ \frac{1}{2\sigma_u^2} \frac{\mu}{\sigma_u} C + \frac{1}{2\sigma^2} \left[\frac{(e_i + \mu)^2}{\sigma^2} - \frac{1}{\lambda \sigma} D_i \left(2\mu + e_i + \frac{\mu}{\lambda^2} \right) - 1 \right] \right\}, \\
U(\sigma_v^2) &= \frac{\partial \ln L}{\partial \sigma_v^2} = \sum_{i=1}^{n} \left\{ \frac{1}{2\sigma^2} \left[\frac{(e_i + \mu)^2}{\sigma^2} + \frac{\lambda}{\sigma} D_i \left(\mu + 2e_i + e_i \lambda^2 \right) - 1 \right] \right\},
\end{aligned} \tag{2.20}$$

em que $C = \dfrac{\phi(c)}{\Phi(c)}$ com $c = \dfrac{\mu}{\sigma_u}$ e $D_i = \dfrac{\phi(d_i)}{\Phi(d_i)}$ com $d_i = \dfrac{\mu}{\sigma\lambda} - \dfrac{e_i \lambda}{\sigma}$.

Note que, com exceção da distribuição exponencial, as distribuições seminormal e normal truncada possuem $\mathrm{Var}(u) \neq \sigma_u^2$, dadas por $\dfrac{\pi - 2}{\pi} \sigma_u^2$ e $\sigma_u^2 \left(1 - \dfrac{\mu}{\sigma_u} a - a^2 \right)$, respectivamente. Ademais, $\mathrm{Var}(u)$ da distribuição normal truncada é função de dois parâmetros, μ e σ_u^2, e $\mathrm{Var}(e)$ se torna não constante se μ for função de covariáveis. Neste caso, assume-se σ_u^2 e σ_v^2 como termos constantes.

Os modelos seminormal e exponencial possuem valor modal da ineficiência técnica igual a zero, com valores crescentes da ineficiência técnica se tornando cada vez menos prováveis (Kumbhakar e Lovell, 2003). Isso implica que a maior parte dos efeitos de ineficiência está na vizinhança de zero e que as medidas de eficiência técnica associadas estarão próximas de um (Coelli et al., 2005).

Os modelos normal truncada e gama permitem uma moda diferente de zero para u, e assim fornecem uma representação um pouco mais flexível do padrão de eficiência técnica dos dados

(Kumbhakar e Lovell, 2003). Infelizmente, esse tipo de flexibilidade vem ao custo da complexidade computacional, na medida em que há mais parâmetros para se estimar. Além disso, se as funções de u_i e de v_i tiverem formas semelhantes, pode ser difícil distinguir os efeitos de ineficiência do ruído (Coelli et al., 2005).

Ainda, diferentes pressupostos distributivos podem dar origem a diferentes predições da eficiência técnica. No entanto, quando classificamos as empresas com base em eficiências técnicas preditas, os ranques costumam ser bastante robustos à escolha distributiva. Em tais casos, o princípio da parcimônia favorece os modelos mais simples - seminormal e exponencial (Coelli et al., 2005).

A maximização de uma função de log-verossimilhança geralmente envolve obter o vetor de primeiras derivadas (gradiente) em relação aos parâmetros desconhecidos e igualá-lo a zero. Infelizmente, no caso das funções de log-verossimilhança acima, estas condições de primeira ordem são altamente não-lineares e não podem ser resolvidas analiticamente. Portanto, um procedimento de otimização iterativa deve ser usado para maximizar essas funções. Isso envolve selecionar valores iniciais para os parâmetros desconhecidos e atualizá-los sistematicamente até que os valores que maximizam a função de log-verossimilhança sejam encontrados. Consequentemente, diferentes implementações podem gerar estimativas um pouco diferentes dos parâmetros do modelo, pois podem usar diferentes algoritmos de otimização iterativa, valores iniciais e/ou critérios de convergência (Coelli et al., 2005).

No caso da distribuição gama, a log-verossimilhança envolverá uma integral intratável que pode ser aproximada numericamente. Greene (1990) e Andrade e Souza (2017) discutem técnicas de aproximação e suas acurácias. Como consequência, o modelo normal/gama não está disponível na maioria das ferramentas estatísticas e econométricas para análise de fronteira estocástica.

Considerações teóricas podem influenciar a escolha da especificação distribucional. Contudo, a escolha da especificação distribucional às vezes é uma questão de conveniência computacional - a estimativa de alguns modelos de fronteira é automatizada em algumas implementações, mas não em outras. Por exemplo, o comando *frontier* e os comandos *sfcross* e *sfpanel*

(Belotti et al., 2013), no *Stata*, podem ser usados para estimar os modelos seminormal, exponencial e normal truncada, enquanto o LIMDEP (Greene, 2012b), que usa uma abordagem de probabilidade simulada, também pode ser usado para estimar o modelo gama.

No entanto, muitas das implementações existentes lidam apenas com a suposição de exogeneidade, não tratando a endogeneidade que pode existir quando uma ou mais das covariáveis da fronteira ou da ineficiência forem correlacionadas com o termo de erro bilateral, v.

Além disso, muitas vezes, os termos de erro podem não ter variância constante. O comando *frontier* do *Stata* permite, para as distribuições seminormal ou exponencial, ajustar modelos com componentes de erro heterocedásticos, condicional a um conjunto de covariáveis. Já para uma distribuição normal truncada, ele permite ajustar um modelo de média condicional, em que a média é modelada como uma função linear de um conjunto de covariáveis.

A maioria dos pacotes existentes no *R* para a análise de fronteira estocástica não permite tal opção, como é o caso do pacote *sfa*, que possibilita especificar as distribuições seminormal, exponencial ou normal truncada para u, mas não permite a inserção de dependência dessas componentes em covariáveis. Tais limitações faz com que seja de interesse a disponibilização de rotinas que possibilitem modelagens sob esses cenários.

2.3 Modelos de fronteira estocástica de produção na presença de endogeneidade

Há muitas aplicações em que as covariáveis, $\boldsymbol{x}_i$, e o erro total, e_i, são não correlacionados no modelo de regressão linear, ou ainda, $\mathrm{E}(e_i|\boldsymbol{x}_i) = 0, i = 1, \ldots, n$. Esta suposição também é chamada de exogeneidade. Quando esse tipo de correlação existe, há endogeneidade. Violações desta suposição podem ocorrer em situações em que há erros de medição, equações simultâneas ou variáveis omitidas. O primeiro, ocorre sempre que as variáveis independentes em um modelo de regressão são medidas com erro. A segunda situação, muitas vezes referida simplesmente como simultaneidade, ocorre sempre que duas ou mais variáveis endógenas são conjuntamente determinadas por um sistema de equações simultâneas (Davidson e MacKinnon, 2004). Enquanto que o terceiro caso, envolve uma grande variedade de situações comuns que envolvem

variáveis que não são observadas, ou por outras razões são omitidas do modelo (Greene, 2012a).

Sem a suposição de que os erros e os regressores são não correlacionados, nenhuma das provas de consistência ou não tendenciosidade do estimador QMO permanecerão válidas, de modo que o estimador QMO perde seu apelo (Greene, 2012a). Para entender porque esse viés ocorre, note que o termo de erro sempre explica parte da variabilidade da variável dependente. No entanto, quando uma variável independente se correlaciona com o termo de erro, o estimador QMO atribui incorretamente uma parte da variação que o termo de erro realmente explica à variável independente. O mesmo ocorre na análise de fronteira estocástica.

Existem duas soluções gerais para o problema de construir um estimador consistente na presença de endogeneidade. Em alguns casos, uma especificação estrutural mais detalhada do modelo pode ser desenvolvida, o que geralmente envolve a especificação de equações adicionais que explicam a correlação entre $\boldsymbol{x}_i$ e e_i de forma a permitir estimar o conjunto completo de parâmetros de interesse. A segunda abordagem é o método de variáveis instrumentais. O método de variáveis instrumentais é desenvolvido em torno da seguinte estratégia de estimação: Suponha que no modelo de regressão linear, K variáveis $\boldsymbol{x}_i$ possam ser correlacionadas com e_i. Suponha também que exista um conjunto L de variáveis $\boldsymbol{z}_i$, de modo que $\boldsymbol{z}_i$ seja correlacionado com $\boldsymbol{x}_i$ (condição de relevância), mas não com e_i (condição de exogeneidade). Não podemos estimar $\boldsymbol{\beta}$ consistentemente usando o estimador usual QMO. Mas, a suposta falta de correlação entre $\boldsymbol{z}_i$ e e_i implica um conjunto de relações que torna possível construir um estimador consistente de $\boldsymbol{\beta}$ usando as relações assumidas entre $\boldsymbol{z}_i$, $\boldsymbol{x}_i$ e e_i (Greene, 2012a).

Em tais situações é conveniente particionar $\boldsymbol{x}$ em dois conjuntos de variáveis, $\boldsymbol{x}_1$, um conjunto de K_1 variáveis exógenas e $\boldsymbol{x}_2$, um conjunto de K_2 variáveis endógenas, com a suposição de que $\boldsymbol{x}_1$ não está correlacionado com e_i e $\boldsymbol{x}_2$ está. Em quase todos os casos, na prática, a endogeneidade é atribuível a um pequeno número de variáveis em $\boldsymbol{x}$. A implicação é que, em tais casos, as K_1 variáveis $\boldsymbol{x}_1$ estarão entre as variáveis instrumentais de $\boldsymbol{Z}$ e as K_2 variáveis remanescentes de $\boldsymbol{Z}$ serão outras variáveis exógenas que não são as mesmas que $\boldsymbol{x}_2$ (variáveis instrumentais externas). A interpretação usual será que essas K_2 variáveis remanescentes, $\boldsymbol{z}_2$, são instrumentos para $\boldsymbol{x}_2$, enquanto as variáveis em $\boldsymbol{x}_1$ são instrumentos para elas mesmas

(Greene, 2012a).

Assim, através do uso de variáveis instrumentais é possível isolar a parte da variável explicativa que não está correlacionada com o erro, tornando possível se obter estimadores consistentes e não viesados dos parâmetros da regressão.

Amsler, Prokhorov e Schmidt (2016) investigam o caso em que um ou mais dos regressores do modelo de fronteira estocástica são endógenos, isto é, quando há correlação entre os regressores e o ruído estatístico ou a ineficiência, no sentido de endogeneidade de equações simultâneas. Eles relatam que isto pode ocorrer quando há *feedback* de qualquer ruído estatístico ou ineficiência na escolha dos regressores, ou quando os regressores influenciam o nível de ineficiência, bem como a fronteira.

Em uma configuração de regressão padrão, a simultaneidade é tratada por um número de procedimentos que são numericamente ou assintoticamente equivalentes. No entanto, procedimentos que são numericamente ou assintoticamente equivalentes no modelo de regressão linear usual podem não ser equivalentes para o modelo de fronteira estocástica.

Como a consistência dos estimadores usuais de fronteira estocástica depende da exogeneidade dos regressores, estes estimadores padrão não lidam com a endogeneidade presente no modelo, que existe se variáveis da fronteira ou da ineficiência estiverem correlacionadas com o termo de erro bilateral, v_i, o que leva a estimativas de parâmetros inconsistentes e, portanto, precisa ser abordada adequadamente. Mutter et al. (2013) explica porque omitir a variável causando endogeneidade não é uma solução viável. Consequentemente, lidar com a questão da endogeneidade na análise de fronteira estocástica é relativamente mais complicado do que nos modelos de regressão padrão, devido à natureza especial do seu termo de erro.

Alguns dos primeiros artigos de fronteira estocástica a lidar com a endogeneidade são Guan et al. (2009), Kutlu (2010) e Tran e Tsionas (2013). Nestes trabalhos, as variáveis podem ser endógenas porque estão correlacionadas com v_i, mas não estão correlacionadas com u_i. Guan et al. (2009) propõem um método de estimação em dois estágios que permite regressores endógenos no modelo. No primeiro estágio de sua metodologia, obtêm-se estimativas consistentes dos parâmetros da fronteira por método dos momentos generalizado e, no segundo estágio, o

resíduo do primeiro estágio é usado como variável dependente para obter estimativas de máxima verossimilhança. Kutlu (2010) descreve um modelo para lidar com a endogeneidade por método de máxima verossimilhança em um e dois estágios, no qual estima a eficiência técnica variável no tempo, na presença de regressores endógenos, através de uma versão modificada do estimador de Battese e Coelli (1992). Tran e Tsionas (2013) propõem uma variação de Kutlu (2010) por método dos momentos generalizado. Contudo, as suposições desses modelos não são suficientes para lidar com a endogeneidade que pode existir devido à correlação entre u_i e v_i.

Tran e Tsionas (2015) e Amsler, Prokhorov e Schmidt (2016) lidam com a endogeneidade por uma abordagem de cópula. Tran e Tsionas (2015) usam uma cópula para permitir modelar diretamente a dependência entre os regressores endógenos e o erro composto. Enquanto que, Amsler, Prokhorov e Schmidt (2016) permite endogeneidade dos regressores em relação ao ruído estatístico e a ineficiência, separadamente. Uma abordagem de cópula permite estruturas de correlação mais gerais ao modelar a endogeneidade. No entanto, este método é computacionalmente intensivo e requer a escolha de uma cópula adequada. Além disso, os modelos propostos em Tran e Tsionas (2015) e Amsler, Prokhorov e Schmidt (2016), bem como em Guan et al. (2009), Kutlu (2010) e Tran e Tsionas (2013), não permitem variáveis ambientais (contextuais) que afetam a ineficiência, o que os tornam menos aplicáveis na tentativa de entender os fatores que afetam a ineficiência.

Griffiths e Hajargasht (2016) e Karakaplan e Kutlu (2015) permitem a endogeneidade em relação aos erros unilateral e bilateral e a correlação entre eles. Além de considerarem variáveis ambientais que afetam a ineficiência. Griffiths e Hajargasht (2016) apresentam um modelo bayesiano de fronteira estocástica onde as ineficiências unilaterais e/ou o termo de erro bilateral podem estar correlacionados com os regressores. Mas seu modelo é muito diferente do modelo proposto por Karakaplan e Kutlu (2015). Karakaplan e Kutlu (2015) sugerem o uso de variáveis instrumentais, em uma metodologia baseada em máxima verossimilhança, para tratar do problema de construir um estimador consistente na presença de endogeneidade devido à correlação entre os termos de erro, permitindo que v_i e u_i dependam de covariáveis que moldam ambas as

distribuições. As estimativas são feitas em um único estágio e lidam com complicações adicionais de modelos de fronteira estocástica que envolvam termos de erro compostos, oferecendo assim uma solução para quando há regressores endógenos.

No geral, um dos principais pontos fortes do modelo de Karakaplan e Kutlu (2015) é que ele é mais fácil de aplicar do que via cópulas ou modelos bayesianos, e é uma generalização direta de um dos modelos de fronteira estocástica mais utilizados, isto é, estimadores do tipo Battese e Coelli (1995). Em seu modelo, Karakaplan e Kutlu (2015) assumem uma regressão linear por variável instrumental, mas a ideia pode ser facilmente generalizada para uma especificação não-linear. Além de que, consideram uma distribuição seminormal para u_i. Karakaplan (2017) disponibiliza o módulo *sfkk* no programa *Stata* para esta especificação de modelo.

Muitos trabalhos empíricos em análise de fronteira estocástica lidam com dados de corte transversal. Contudo, se disponível, um painel (observações repetidas sobre cada produtor) geralmente contém mais informações do que um único corte transversal. Consequentemente, é de se esperar que o acesso a dados em painel permita que algumas das premissas distribucionais usadas com dados de corte transversal sejam relaxadas ou resultem em estimativas de eficiência técnica com propriedades estatísticas mais desejáveis (Kumbhakar e Lovell, 2003).

Os dados do painel geralmente permitem (Coelli et al., 2005):

(i) Relaxar algumas das premissas distributivas necessárias para separar os efeitos de ineficiência e ruído.

(ii) Obter predições consistentes das eficiências técnicas.

(iii) Investigar mudanças nas eficiências técnicas (assim como a tecnologia de produção subjacente) ao longo do tempo.

Consequentemente, observações repetidas em uma amostra de produtores podem servir como um substituto para premissas distribucionais e para a suposição de independência. Algumas limitações, tais como, problemas relacionados a heterocedasticidade em u_i, v_i ou ambos, ou ainda, a suposição de que $u_i \perp \boldsymbol{x}_i$ - embora seja fácil imaginar que a ineficiência técnica possa estar correlacionada com os vetores de insumos selecionados pelos produtores - são evitáveis se tivermos acesso aos dados do painel. Por exemplo, o modelo de efeitos fixos com

eficiência técnica invariável no tempo assume que os u_i são fixos, mas permite que eles sejam correlacionados com os regressores (Coelli et al., 2005).

Portanto, uma vez que a adição de mais observações sobre cada produtor gera informações não fornecidas pela adição de mais produtores a um corte transversal, a eficiência técnica de cada produtor na amostra pode ser estimada consistentemente quando $T \to +\infty$, sendo T o número de observações sobre cada produtor. No entanto, esse benefício final de ter acesso aos dados do painel pode ser exagerado, pois muitos painéis são relativamente curtos (Kumbhakar e Lovell, 2003).

2.4 Considerações finais

Este capítulo apresentou alguns modelos usados para modelar a ineficiência técnica da unidade de produção, na presença ou não de variáveis endógenas, na análise de fronteira estocástica - especificamente, os modelos seminormal, exponencial e normal truncada. Além disso, abordou as consequências da presença de endogeneidade e sinteticamente apresentou algumas técnicas que podem ser utilizadas nesse caso.

O próximo capítulo pretende explorar o método de estimação por máxima verossimilhança em um estágio proposto por Karakaplan e Kutlu (2015), para lidar com a endogeneidade. O intuito é estender sua técnica para além da distribuição seminormal.

Capítulo 3

Método de máxima verossimilhança em um estágio

3.1 Considerações iniciais

O objetivo deste capítulo é abordar um método de estimação de parâmetros do modelo de fronteira estocástica de produção na presença de endogeneidade via máxima verossimilhança em um estágio (abordagem MVIC), apresentada por Karakaplan e Kutlu (2015), estendendo a metodologia proposta originalmente para a distribuição seminormal, para as distribuições exponencial e normal truncada do termo de ineficiência.

3.2 Modelo

Considere o seguinte modelo de fronteira estocástica de produção com covariáveis endógenas, proposto por Karakaplan e Kutlu (2015):

$$\begin{aligned} y_i &= x_{1i}^{\top}\beta + v_i - u_i, \\ x_i &= Z_i\delta + \varepsilon_i, \\ \begin{bmatrix} \tilde{\varepsilon}_i \\ v_i \end{bmatrix} &\equiv \begin{bmatrix} \Omega^{-\frac{1}{2}}\varepsilon_i \\ v_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} I_p & \sigma_{vi}\rho \\ \sigma_{vi}\rho^{\top} & \sigma_{vi}^2 \end{bmatrix} \right). \end{aligned} \tag{3.1}$$

No modelo da Eq. (3.1), y_i é o logaritmo da produção do i-ésimo produtor; x_{1i} é um vetor de variáveis exógenas e endógenas; x_i é um vetor $p \times 1$ de todas as variáveis endógenas (excluindo y_i); $Z_i = I_p \otimes z_i^\top$, sendo z_i um vetor $q \times 1$ de todas as variáveis exógenas e instrumentais; v_i e ε_i são termos de erros bilaterais; $u_i \geq 0$ é um termo de erro unilateral capturando a ineficiência; Ω é a matriz de variância-covariância de ε_i; σ_{vi}^2 é a variância de v_i; ρ é um vetor que representa a correlação entre $\tilde{\varepsilon}_i$ e v_i. Nesta estrutura, uma variável é endógena se não for independente do termo de erro bilateral, v_i.

As especificações de modelo propostas por Karakaplan e Kutlu (2015) fornecem uma metodologia para lidar com questões de endogeneidade em modelos de fronteira estocástica em um cenário mais geral. O modelo considera heteroscedasticidade nos componentes do termo de erro composto, ao permitir que u_i e v_i sejam dependentes através de covariáveis que moldam ambas as distribuições.

Seja x_{2i} um vetor de variáveis exógenas e endógenas. Assuma que o termo de ineficiência, u_i, é uma função de x_{2i} e de um componente aleatório, u_i^*, específico de cada produtor, isto é, $u_i = \sigma_u(x_{2i}; \varphi_u)u_i^*$, em que $\sigma_{ui} = \sigma_u(x_{2i}; \varphi_u) > 0$ e $u_i^* \geq 0$. Os termos σ_{ui} e u_i^* são independentes de v_i e ε_i, condicional em x_i e z_i. Portanto, u_i não é independente de x_i, mas u_i e v_i são condicionalmente independentes dado x_i e z_i. Do mesmo modo, u_i e ε_i são condicionalmente independentes dado x_i e z_i. Em geral, $\mathrm{Cov}(u_i, \varepsilon_i) \neq 0$. Esta é uma das características importantes deste modelo, pois os modelos convencionais de fronteira estocástica não permitem tais correlações.

Pela decomposição de Cholesky da matriz de variância-covariância de $(\tilde{\varepsilon}_i^\top, v_i)^\top$, é possível representar $(\tilde{\varepsilon}_i^\top, v_i)^\top$ como

$$\begin{bmatrix} \tilde{\varepsilon}_i \\ v_i \end{bmatrix} = \begin{bmatrix} I_p & 0 \\ \sigma_{vi}\rho^\top & \sigma_{vi}\sqrt{1-\rho^\top\rho} \end{bmatrix} \begin{bmatrix} \tilde{\varepsilon}_i \\ \tilde{w}_i \end{bmatrix}, \tag{3.2}$$

em que $\tilde{\varepsilon}_i$ e $\tilde{w}_i \sim \mathcal{N}(0,1)$ são independentes. Assim, a equação de fronteira pode ser escrita como

$$\begin{aligned} y_i &= x_{1i}^\top \beta + \sigma_{vi} \rho^\top \tilde{\varepsilon}_i + w_i - u_i \\ &= x_{1i}^\top \beta + \frac{\sigma_{wi}}{\sigma_{cw}} \eta^\top (x_i - Z_i \delta) + e_i, \end{aligned} \tag{3.3}$$

na qual $e_i = w_i - u_i$, $w_i = \sigma_{vi}\sqrt{1-\rho^\top\rho}\tilde{w}_i = \sigma_{wi}\tilde{w}_i$, $\sigma_{wi} = \sigma_{cw}\sigma_w(.;\varphi_w)$ é separável de modo que $\sigma_{cw} > 0$ é uma função do termo constante, $\sigma_w(.;\varphi_w)$ é uma função de todas as variáveis afetando σ_{wi}, exceto o termo constante, e $\eta = \sigma_{cw}\Omega^{-\frac{1}{2}}\rho/\sqrt{1-\rho^\top\rho}$. Então, quando não há heterocedasticidade em w_i, $\sigma_{wi} = \sigma_{cw}$, tal que

$$y_i = x_{1i}^\top \beta + \eta^\top (x_i - Z_i \delta) + e_i. \tag{3.4}$$

O termo e_i é condicionalmente independente dos regressores dado x_i e z_i e é possível diretamente assumir que a distribuição condicional de v_i dado x_i (e variáveis exógenas) é uma distribuição normal com média $(\sigma_{wi}/\sigma_{cw})\eta^\top(x_i - Z_i\delta)$. Esta abordagem é comumente usada para resolver o problema de construir um estimador consistente, na presença de endogeneidade, em modelos com não-linearidade intrínseca, tal como este modelo, no qual $(\sigma_{wi}/\sigma_{cw})\eta^\top(x_i - Z_i\delta)$ é um termo de correção de viés. Portanto, esta abordagem trata a endogeneidade como um problema de variável omitida.

Karakaplan e Kutlu (2015) assumem que: $u_i^* \sim \mathcal{N}^+(0,1)$, $\sigma_{ui}^2 = \exp(x_{2i}^\top\varphi_u)$, $\sigma_{wi}^2 = \exp(x_{3i}^\top\varphi_w)$, $\sigma_{cw}^2 = \exp(\varphi_{cw})$, em que $\varphi = (\varphi_u^\top, \varphi_w^\top)^\top$ é o vetor de parâmetros capturando a heterocedasticidade e x_{3i} é um vetor de variáveis exógenas e endógenas que podem compartilhar as mesmas variáveis com x_{1i} e x_{2i}. φ_{cw} é o coeficiente do termo constante para $x_{3i}^\top\varphi_w$. Implicando que $u_i \sim \mathcal{N}^+(0, \sigma_{ui}^2)$.

É possível escolher outras funções de ligação para σ_{ui}^2 e σ_{wi}^2, tal como a função quadrática, além de outra distribuição para o componente de ineficiência, por exemplo, a distribuição exponencial ou a normal truncada. Se $u_i^* \sim \text{Exp}(1)$ então $u_i \sim \text{Exp}(\sigma_{ui})$ e se $u_i^* \sim \mathcal{N}^+(\mu_i, 1)$ então $u_i \sim \mathcal{N}^+(\mu_i, \sigma_u^2)$, com $\mu_i = x_{2i}^\top\tau$ sendo a média da distribuição normal truncada, na qual τ é o vetor de parâmetros capturando a heterocedasticidade.[1].

[1]No caso da distribuição normal truncada, assume-se que apenas μ_i é função de covariáveis, enquanto que σ_{ui}^2 e σ_{wi}^2 são termos constantes.

3.3 Log-verossimilhança

Seja $y = (y_1, \ldots, y_n)^\top$ o vetor de observações da variável dependente, $x = (x_1^\top, \ldots, x_n^\top)^\top$ a matriz de variáveis endógenas no modelo (isto é, os elementos de x são os x_i's definidos anteriormente) e $\theta = (\beta^\top, \eta^\top, \varphi^\top, \delta^\top, \tau^\top)^\top$ o vetor de coeficientes. A log-verossimilhança do modelo de fronteira estocástica de produção pode ser decomposta em duas partes, $\ln L(\theta) = \ln L_{y|x}(\theta) + \ln L_x(\theta)$, na qual

$$
\begin{aligned}
\ln L_{y|x}(\theta) &= \text{distribuição normal/seminormal, normal/exponencial ou normal/normal truncada} \\
\ln L_x(\theta) &= \sum_{i=1}^{n} \left\{ \frac{-p\ln(2\pi) - \ln(|\Omega|) - \varepsilon_i^\top \Omega^{-1} \varepsilon_i}{2} \right\} \\
e_i &= y_i - x_{1i}^\top \beta - \frac{\sigma_{wi}}{\sigma_{cw}} \eta^\top (x_i - Z_i \delta) \\
\varepsilon_i &= x_i - Z_i \delta \\
\sigma_i^2 &= \sigma_{ui}^2 + \sigma_{wi}^2 \\
\lambda_i &= \frac{\sigma_{ui}}{\sigma_{wi}} \\
\mu_i &\neq 0 \text{ se e somente se } u_i \sim \mathcal{N}^+(\mu_i, \sigma_u^2),
\end{aligned}
\tag{3.5}
$$

Portanto, x segue uma distribuição normal multivariada se o número de variáveis endógenas é maior que um e normal univariada, caso contrário. Enquanto que $y|x$ segue uma distribuição normal/seminormal (Eq. 2.7), normal/exponencial (Eq. 2.13) ou normal/normal truncada (Eq. 2.18). $\ln L_x(\theta)$ é adicionada a $\ln L_{y|x}(\theta)$ e e_i é ajustado pelo fator $(\sigma_{wi}/\sigma_{cw})\eta^\top(x_i - Z_i\delta)$, que resolve o problema de estimativas de parâmetros inconsistentes devido a regressores endógenos em x_{1i} e devido às variáveis endógenas em x_{2i}. Ainda é possível testar quanto à presença de endogeneidade, ao testar a hipótese nula de que $\eta = 0$. Mais informações em Karakaplan e Kutlu (2015).

3.4 Gradientes

Para os gradientes a seguir assume-se função de ligação exponencial para σ_{ui}^2 e σ_{wi}^2 e que $\mu_i \in \mathbb{R}$.

Da Eq. (3.5), ao assumir distribuição seminormal para u_i, o gradiente é dado por

$$\begin{aligned}
U(\delta) &= \sum_{i=1}^{n} Z_i^\top \varepsilon_i \Omega^{-1} - \sum_{i=1}^{n} Z_i^\top \left\{ \frac{\sigma_{wi}}{\sigma_{cw}} \left[\frac{e_i}{\sigma_i^2} + \frac{\lambda_i}{\sigma_i} A_i \right] \right\} \eta^\top, \\
U(\beta) &= \sum_{i=1}^{n} x_{1i}^\top \left\{ \frac{e_i}{\sigma_i^2} + \frac{\lambda_i}{\sigma_i} A_i \right\}, \\
U(\eta) &= \sum_{i=1}^{n} \varepsilon_i^\top \left\{ \frac{\sigma_{wi}}{\sigma_{cw}} \left[\frac{e_i}{\sigma_i^2} + \frac{\lambda_i}{\sigma_i} A_i \right] \right\}, \\
U(\varphi_u) &= \sum_{i=1}^{n} x_{2i}^\top \left\{ \frac{1}{2\sigma_i^2} \left[\frac{e_i^2}{\sigma_i^2} - \frac{e_i}{\lambda_i \sigma_i} A_i - 1 \right] \right\} \sigma_{ui}^2, \\
U(\varphi_w) &= \sum_{i=1}^{n} x_{3i}^\top \left\{ \frac{1}{2\sigma_i^2} \left[\frac{e_i^2}{\sigma_i^2} + \frac{e_i \lambda_i}{\sigma_i} A_i \left(2 + \lambda_i^2 \right) - 1 \right] \right\} \sigma_{wi}^2 + \\
&\quad + \sum_{i=1}^{n} \tilde{x}_{3i}^\top \left\{ \frac{1}{2} \eta^\top \varepsilon_i \frac{\sigma_{wi}}{\sigma_{cw}} \left[\frac{e_i}{\sigma_i^2} + \frac{\lambda_i}{\sigma_i} A_i \right] \right\},
\end{aligned} \tag{3.6}$$

em que $A_i = \dfrac{\phi(a_i)}{\Phi(a_i)}$ com $a_i = -\dfrac{e_i \lambda_i}{\sigma_i}$. $\tilde{x}_{3i}$ é x_{3i} exceto pelo componente de intercepto nulo.

Da Eq. (3.5), ao assumir distribuição exponencial para u_i, o gradiente é da forma

$$\begin{aligned}
U(\delta) &= \sum_{i=1}^{n} Z_i^\top \varepsilon_i \Omega^{-1} - \sum_{i=1}^{n} Z_i^\top \left\{ \frac{\sigma_{wi}}{\sigma_{cw}} \left[\frac{1}{\sigma_{wi}} B_i - \frac{1}{\sigma_{ui}} \right] \right\} \eta^\top, \\
U(\eta) &= \sum_{i=1}^{n} \varepsilon_i^\top \left\{ \frac{\sigma_{wi}}{\sigma_{cw}} \left[\frac{1}{\sigma_{wi}} B_i - \frac{1}{\sigma_{ui}} \right] \right\}, \\
U(\beta) &= \sum_{i=1}^{n} x_{1i}^\top \left\{ \frac{1}{\sigma_{wi}} B_i - \frac{1}{\sigma_{ui}} \right\}, \\
U(\varphi_u) &= \sum_{i=1}^{n} x_{2i}^\top \left\{ \frac{1}{2\sigma_{ui}^2} \left[\frac{\sigma_{wi}}{\sigma_{ui}} B_i - \frac{\sigma_{wi}^2}{\sigma_{ui}^2} - \frac{e_i}{\sigma_{ui}} - 1 \right] \right\} \sigma_{ui}^2, \\
U(\varphi_w) &= \sum_{i=1}^{n} x_{3i}^\top \left\{ \frac{1}{2\sigma_{ui}^2} + \frac{1}{2\sigma_{wi}} B_i \left[\frac{e_i}{\sigma_{wi}^2} - \frac{1}{\sigma_{ui}} \right] \right\} \sigma_{wi}^2 + \\
&\quad + \sum_{i=1}^{n} \tilde{x}_{3i}^\top \left\{ \frac{1}{2} \eta^\top \varepsilon_i \frac{\sigma_{wi}}{\sigma_{cw}} \left[\frac{1}{\sigma_{wi}} B_i - \frac{1}{\sigma_{ui}} \right] \right\},
\end{aligned} \tag{3.7}$$

em que $B_i = \dfrac{\phi(b_i)}{\Phi(b_i)}$ com $b_i = \dfrac{-e_i - \sigma_{wi}^2/\sigma_{ui}}{\sigma_{wi}}$. $\tilde{x}_{3i}$ é x_{3i} exceto pelo componente de intercepto

nulo.

Da Eq. (3.5), ao assumir distribuição normal truncada para u_i, o gradiente é dado por

$$\begin{aligned}
U(\delta) &= \sum_{i=1}^{n} Z_i^\top \varepsilon_i \Omega^{-1} - \sum_{i=1}^{n} Z_i^\top \left\{ \frac{\sigma_w}{\sigma_{cw}} \left[\frac{e_i + \mu_i}{\sigma^2} + \frac{\lambda}{\sigma} D_i \right] \right\} \eta^\top, \\
U(\eta) &= \sum_{i=1}^{n} \varepsilon_i^\top \left\{ \frac{\sigma_w}{\sigma_{cw}} \left[\frac{e_i + \mu_i}{\sigma^2} + \frac{\lambda}{\sigma} D_i \right] \right\}, \\
U(\beta) &= \sum_{i=1}^{n} x_{1i}^\top \left\{ \frac{e_i + \mu_i}{\sigma^2} + \frac{\lambda}{\sigma} D_i \right\}, \\
U(\tau) &= \sum_{i=1}^{n} x_{2i}^\top \left\{ \frac{1}{\lambda\sigma} D_i - \frac{\sqrt{\lambda^{-2}+1}}{\sigma} C_i - \frac{e_i + \mu_i}{\sigma^2} \right\}, \\
U(\varphi_u) &= \sum_{i=1}^{n} \left\{ \frac{1}{2\sigma_u^2} \frac{\mu_i}{\sigma_u} C_i + \frac{1}{2\sigma^2} \left[\frac{(e_i + \mu_i)^2}{\sigma^2} - \frac{1}{\lambda\sigma} D_i \left(2\mu_i + e_i + \frac{\mu_i}{\lambda^2} \right) - 1 \right] \right\} \sigma_u^2, \\
U(\varphi_w) &= \sum_{i=1}^{n} \left\{ \frac{1}{2\sigma^2} \left[\frac{(e_i + \mu_i)^2}{\sigma^2} + \frac{\lambda}{\sigma} D_i \left(\mu_i + 2e_i + e_i\lambda^2 \right) - 1 \right] \right\} \sigma_w^2,
\end{aligned} \tag{3.8}$$

em que $C_i = \dfrac{\phi(c_i)}{\Phi(c_i)}$ com $c_i = \dfrac{\mu_i}{\sigma_u}$ e $D_i = \dfrac{\phi(d_i)}{\Phi(d_i)}$ com $d_i = \dfrac{\mu_i}{\sigma\lambda} - \dfrac{e_i\lambda}{\sigma}$.

A estimação do vetor θ é feita em um único estágio (simultaneamente), caracterizando uma abordagem de máxima verossimilhança de informação completa (MVIC).

3.5 Considerações finais

Este capítulo apresentou uma abordagem de máxima verossimilhança em um estágio na presença de endogeneidade, assumindo distribuição seminormal, exponencial ou normal truncada para o termo de ineficiência e função de ligação exponencial para σ_{ui}^2 e σ_{wi}^2. Além disso, o gradiente analítico dos três modelos considerados foi apresentado.

O próximo capítulo pretende explorar modelos para lidar com a endogeneidade por método de máxima verossimilhança em dois estágios, com correção da variância proposta por Murphy e Topel (1985).

Capítulo 4

Método de máxima verossimilhança em dois estágios

4.1 Considerações iniciais

O objetivo deste capítulo é abordar um método de estimação de parâmetros do modelo de fronteira estocástica de produção na presença de endogeneidade via máxima verossimilhança em dois estágios (abordagem MVIL), com correção da matriz de variância-covariância proposta por Murphy e Topel (1985).

4.2 Modelo

Problemas de estimação em dois estágios são aqueles em que elementos de um modelo são embutidos em outro e caracterizados por

$$\begin{aligned} &\text{Modelo 1: } E\{y_1|x_1,\theta_1\}, \\ &\text{Modelo 2: } E\{y_2|x_2,\theta_2,\mathrm{E}(y_1|x_1,\theta_1)\}. \end{aligned} \tag{4.1}$$

Há dois vetores de parâmetros a serem estimados. O primeiro vetor, θ_1, aparece em ambos os modelos, mas o segundo vetor, θ_2, aparece apenas no segundo modelo. Embora θ_1 e θ_2 possam ser estimados em conjunto, a adaptação dos modelos usando um procedimento de dois

estágios é computacionalmente mais fácil. Nesta abordagem, o modelo 1 é ajustado primeiro, pois não envolve o segundo vetor de parâmetros. Em seguida o modelo 2 é ajustado de acordo com os resultados do primeiro estágio.

Em tais situações, duas abordagens podem ser usadas para se obter as estimativas dos parâmetros envolvidos: máxima verossimilhança de informação completa (MVIC) e máxima verossimilhança de informação limitada (MVIL). A abordagem MVIC é aquela em que é especificada a distribuição conjunta e maximizada a função de log-verossimilhança conjunta, dada por $\sum_{i=1}^{n} \ln f(y_{1i}, y_{2i}|x_{1i}, x_{2i}, \theta_1, \theta_2)$. Portanto, uma abordagem MVIC é usada em um estágio e a estimação de θ_1 e θ_2 é feita simultaneamente. Na abordagem MVIL, primeiramente é estimado θ_1 e condicional a estas estimativas é estimado o segundo vetor de parâmetros, θ_2. Assim, a função de log-verossimilhança condicional a ser maximizada é dada por $\sum_{i=1}^{n} \ln f(y_{2i}|x_{2i}, \theta_2, (x_{1i}, \hat{\theta}_1))$ (Greene, 2012a).

Há pelo menos duas razões pelas quais alguém pode usar uma estimação em dois estágios. Primeiro, pode ser direto formular as duas log-verossimilhanças separadas, mas muito complicado derivar a distribuição conjunta. Essa situação surge com frequência quando as duas variáveis que estão sendo modificadas são de diferentes tipos de populações, como uma discreta e outra contínua (o que é um caso muito comum nesta estrutura). A segunda razão é que maximizar as log-verossimilhanças separadas pode ser bastante simples, mas maximizar a log-verossimilhança conjunta pode ser numericamente complicado ou difícil (Greene, 2012a).

Uma desvantagem em relação ao método de um estágio, como o apresentado no capítulo anterior, é que embora a abordagem de estimação em dois estágios leve a uma estimativa consistente de θ_2, a matriz de variância-covariância estimada para o modelo 2 precisa ser ajustada para levar em conta a variabilidade em $\hat{\theta}_1$, desde que $\hat{\theta}_1$ é uma estimativa de θ_1 em vez de seu valor real. Portanto, o estimador de dois estágios fornece erros padrão incorretos e inconsistentes, sendo necessária uma correção desses erros. Para tanto, uma abordagem analítica é possível, como a proposta por Murphy e Topel (1985).

4.3 Correção da variância de Murphy e Topel (1985)

A metodologia proposta por Murphy e Topel (1985) para a correção da variância é usada para abordar o fato de que um ou mais dos regressores foram gerados via $(x_{1i}, \hat{\theta}_1)$. Os autores descrevem uma fórmula geral de um estimador da variância válido para θ_2 em um modelo de estimação por máxima verossimilhança em dois estágios.

Se as condições de regularidade padrões são atendidas para ambas as funções, então o estimador de máxima verossimilhança em dois estágios de θ_2 é consistente e assintoticamente normalmente distribuído com matriz de variância-covariância (Greene, 2012a)

$$\mathbf{V_2^*} = \frac{1}{n}\big[\mathbf{V_2} + \mathbf{V_2}(\mathbf{C}\mathbf{V_1}\mathbf{C}^\top - \mathbf{R}\mathbf{V_1}\mathbf{C}^\top - \mathbf{C}\mathbf{V_1}\mathbf{R}^\top)\mathbf{V_2}\big], \tag{4.2}$$

em que

$$\begin{aligned}
\mathbf{V_1} &= \text{matriz } (q \times q) \text{ de variância assintótica de } \hat{\theta}_1 \text{ baseada em } \ln L_1(\theta_1), \\
\mathbf{V_2} &= \text{matriz } (p \times p) \text{ de variância assintótica de } \hat{\theta}_2 \text{ baseada em } \ln L_2(\theta_2|\theta_1), \\
\mathbf{C} &= \text{matriz } (p \times q) \text{ dada por } E\left[\frac{1}{n}\left(\frac{\partial \ln L_2}{\partial \theta_2}\right)\left(\frac{\partial \ln L_2}{\partial \theta_1^\top}\right)\right], \\
\mathbf{R} &= \text{matriz } (p \times q) \text{ dada por } E\left[\frac{1}{n}\left(\frac{\partial \ln L_2}{\partial \theta_2}\right)\left(\frac{\partial \ln L_1}{\partial \theta_1^\top}\right)\right].
\end{aligned} \tag{4.3}$$

As matrizes $\mathbf{V_1}$ e $\mathbf{V_2}$ são estimadas pelas matrizes de variância-covariância não corrigidas dos modelos 1 e 2, respectivamente, tipicamente pelo estimador *BHHH* (Eq. 4.5) ou pelas matrizes inversas de segundas derivadas negativas (Eq. 4.6). Já as matrizes $\mathbf{C}$ e $\mathbf{R}$ são obtidas somando as observações individuais nos produtos cruzados das derivadas. Assume-se a existência de uma log-verossimilhança para o primeiro modelo, $\ln L_1(\theta_1)$, e uma log-verossimilhança condicional para o segundo modelo (primário) de interesse, $\ln L_2(\theta_2|\theta_1)$. As matrizes componentes do estimador de Murphy-Topel são estimadas pela avaliação das fórmulas nas estimativas de máxima verossimilhança de $\hat{\theta}_1$ e $\hat{\theta}_2$. Assim,

$$\hat{\mathbf{V}}_\mathbf{2}^* = \frac{1}{n}\Big[\hat{\mathbf{V}}_\mathbf{2} + \hat{\mathbf{V}}_\mathbf{2}(\hat{\mathbf{C}}\hat{\mathbf{V}}_\mathbf{1}\hat{\mathbf{C}}^\top - \hat{\mathbf{R}}\hat{\mathbf{V}}_\mathbf{1}\hat{\mathbf{C}}^\top - \hat{\mathbf{C}}\hat{\mathbf{V}}_\mathbf{1}\hat{\mathbf{R}}^\top)\hat{\mathbf{V}}_\mathbf{2}\Big], \tag{4.4}$$

em que

$$\begin{aligned}
\hat{\mathbf{V}}_{\mathbf{1}} &= \left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{1i}}{\partial \hat{\theta}_1}\right)\left(\frac{\partial \ln f_{1i}}{\partial \hat{\theta}_1^\top}\right)\right]^{-1} \qquad \hat{\mathbf{C}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{2i}}{\partial \hat{\theta}_2}\right)\left(\frac{\partial \ln f_{2i}}{\partial \hat{\theta}_1^\top}\right) \\
\hat{\mathbf{V}}_{\mathbf{2}} &= \left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{2i}}{\partial \hat{\theta}_2}\right)\left(\frac{\partial \ln f_{2i}}{\partial \hat{\theta}_2^\top}\right)\right]^{-1} \qquad \hat{\mathbf{R}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{2i}}{\partial \hat{\theta}_2}\right)\left(\frac{\partial \ln f_{1i}}{\partial \hat{\theta}_1^\top}\right)
\end{aligned} \tag{4.5}$$

ou ainda,[1]

$$\begin{aligned}
\hat{\mathbf{V}}_{\mathbf{1}} &= \left[-\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial^2 \ln f_{1i}}{\partial \hat{\theta}_1 \partial \hat{\theta}_1^\top}\right)\right]^{-1} \qquad \hat{\mathbf{C}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{2i}}{\partial \hat{\theta}_2}\right)\left(\frac{\partial \ln f_{2i}}{\partial \hat{\theta}_1^\top}\right) \\
\hat{\mathbf{V}}_{\mathbf{2}} &= \left[-\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial^2 \ln f_{2i}}{\partial \hat{\theta}_2 \partial \hat{\theta}_2^\top}\right)\right]^{-1} \qquad \hat{\mathbf{R}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{2i}}{\partial \hat{\theta}_2}\right)\left(\frac{\partial \ln f_{1i}}{\partial \hat{\theta}_1^\top}\right)
\end{aligned} \tag{4.6}$$

4.4 Log-verossimilhança

Seja $y = (y_1, \ldots, y_n)^\top$ o vetor de observações da variável dependente, $x = (x_1^\top, \ldots, x_n^\top)^\top$ a matriz de variáveis endógenas no modelo e $\theta = (\beta^\top, \eta^\top, \varphi^\top, \delta^\top, \tau^\top)^\top$ o vetor de coeficientes. A log-verossimilhança do modelo pode ser decomposta em duas partes, na qual

$$\begin{aligned}
&\ln L_2(\theta_2|\theta_1) = \text{distribuição normal/seminormal, normal/exponencial ou normal/normal truncada} \\
&\ln L_1(\theta_1) = \sum_{i=1}^{n}\left\{\frac{-p\ln(2\pi) - \ln(|\Omega|) - \varepsilon_i^\top \Omega^{-1}\varepsilon_i}{2}\right\} \\
&\qquad e_i = y_i - x_{1i}^\top \beta - \hat{\varepsilon}_i^\top \eta \\
&\qquad \varepsilon_i = x_i - Z_i\delta \\
&\qquad \sigma_i^2 = \sigma_{ui}^2 + \sigma_{wi}^2 \\
&\qquad \lambda_i = \frac{\sigma_{ui}}{\sigma_{wi}} \\
&\qquad \mu_i \neq 0 \text{ se e somente se } u_i \sim \mathcal{N}^+(\mu_i, \sigma_u^2),
\end{aligned} \tag{4.7}$$

Assim, no primeiro estágio, $\ln L_x(\theta) = \ln L_1(\theta_1)$ é maximizada em relação aos parâmetros

[1] Neste trabalho, os estimadores da Eq. (4.6) são utilizados para se obter $\hat{\mathbf{V}}_{\mathbf{2}}^*$.

relevantes. No segundo estágio, condicional aos parâmetros estimados no primeiro estágio, $\ln L_{y|x}(\theta) = \ln L_2(\theta_2|\theta_1)$ é maximizada. Nesta metodologia, o modelo do segundo estágio é

$$y_i = x_{1i}^\top \beta + \eta^\top \hat{\varepsilon}_i + e_i, \tag{4.8}$$

no qual $e_i = w_i - u_i$ e $\hat{\varepsilon}_i$ são as estimativas dos resíduos do primeiro estágio obtidas via QMO por meio da equação $\hat{\varepsilon}_i = x_i - Z_i\hat{\delta}$. Os coeficientes dos termos $\hat{\varepsilon}_i$ podem ser testados quanto à presença de endogeneidade, ou seja, é possível testar a hipótese nula de que $\eta = 0$. Nesta estrutura, uma variável é endógena se não for independente de v_i.

Assim como no modelo proposto por Karakaplan e Kutlu (2015), nesta abordagem, o modelo 1 segue uma distribuição normal multivariada se o número de variáveis endógenas é maior que um e normal univariada, caso contrário, e o modelo 2 é especificado como normal/seminormal (Eq. 2.7), normal/exponencial (Eq. 2.13) ou normal/normal truncada (Eq. 2.18).

No modelo 1, $x_i = Z_i\delta + \varepsilon_i$ é um vetor $p \times 1$ de todas as variáveis endógenas (excluindo y_i) e δ é um vetor de parâmetros desconhecidos associados a $Z_i = I_p \otimes z_i^\top$, sendo z_i um vetor $q \times 1$ de todas as variáveis exógenas e instrumentais.

Para o modelo 2, assume-se que $\lambda_i = \sigma_{ui}/\sigma_{wi}$, $\sigma_i^2 = \sigma_{ui}^2 + \sigma_{wi}^2$ e $e_i = y_i - x_{1i}^\top\beta - \eta^\top\hat{\varepsilon}_i$, em que y_i é um vetor de observações do logaritmo da produção dos i produtores, x_{1i} é um vetor de variáveis exógenas e endógenas, $\mu_i = x_{2i}^\top\tau$ é a média da distribuição normal truncada, $\sigma_{ui}^2 = \exp(x_{2i}^\top\varphi_u)$ e $\sigma_{wi}^2 = \exp(x_{3i}^\top\varphi_w)$, nos quais μ_i, σ_{ui}^2 e σ_{wi}^2 podem ser funções de covariáveis[2].

A estimação em dois estágios, além de ser uma alternativa mais fácil de implementar, pode ser estendida para acomodar modelos de regressão linear ou não linear por variável instrumental. Neste trabalho, primeiramente, ajusta-se a regressão por variável instrumental e calcula-se os termos $\hat{\varepsilon}_i$ e, em seguida, executa-se o modelo de fronteira estocástica, dado na Eq. (4.8).

Esse processo de estimação pela abordagem MVIL não produz os mesmos resultados que a abordagem MVIC, mostrada no Capítulo 3, e sua matriz de variância-covariância do segundo estágio deve ser ajustada. No entanto, ele apresenta menos problemas de convergência.

[2]Com exceção da distribuição normal truncada, que permite apenas μ_i como função de covariáveis, enquanto que σ_{ui}^2 e σ_{wi}^2 são termos constantes.

4.5 Gradientes

O gradiente da Eq. (4.7), ao assumir distribuição seminormal para u_i, é dado por

$$\begin{aligned}
U_1(\delta) &= -2\sum_{i=1}^{n} Z_i^\top \varepsilon_i, \\
U_2(\delta) &= -\sum_{i=1}^{n} Z_i^\top \left\{ \frac{e_i}{\sigma_i^2} + \frac{\lambda_i}{\sigma_i} A_i \right\} \eta^\top, \\
U_2(\eta) &= \sum_{i=1}^{n} \hat{\varepsilon}_i^\top \left\{ \frac{e_i}{\sigma_i^2} + \frac{\lambda_i}{\sigma_i} A_i \right\}, \\
U_2(\beta) &= \sum_{i=1}^{n} x_{1i}^\top \left\{ \frac{e_i}{\sigma_i^2} + \frac{\lambda_i}{\sigma_i} A_i \right\}, \\
U_2(\varphi_u) &= \sum_{i=1}^{n} x_{2i}^\top \left\{ \frac{1}{2\sigma_i^2} \left[\frac{e_i^2}{\sigma_i^2} - \frac{e_i}{\lambda_i \sigma_i} A_i - 1 \right] \right\} \sigma_{ui}^2, \\
U_2(\varphi_w) &= \sum_{i=1}^{n} x_{3i}^\top \left\{ \frac{1}{2\sigma_i^2} \left[\frac{e_i^2}{\sigma_i^2} + \frac{e_i \lambda_i}{\sigma_i} (2 + \lambda_i^2) A_i - 1 \right] \right\} \sigma_{wi}^2,
\end{aligned} \tag{4.9}$$

em que $A_i = \dfrac{\phi(a_i)}{\Phi(a_i)}$ com $a_i = -\dfrac{e_i \lambda_i}{\sigma_i}$.

O gradiente da Eq. (4.7), ao assumir distribuição exponencial para u_i, é dado por

$$\begin{aligned}
U_1(\delta) &= -2\sum_{i=1}^{n} Z_i^\top \varepsilon_i, \\
U_2(\delta) &= -\sum_{i=1}^{n} Z_i^\top \left\{ \frac{1}{\sigma_{wi}} B_i - \frac{1}{\sigma_{ui}} \right\} \eta^\top, \\
U_2(\eta) &= \sum_{i=1}^{n} \hat{\varepsilon}_i^\top \left\{ \frac{1}{\sigma_{wi}} B_i - \frac{1}{\sigma_{ui}} \right\}, \\
U_2(\beta) &= \sum_{i=1}^{n} x_{1i}^\top \left\{ \frac{1}{\sigma_{wi}} B_i - \frac{1}{\sigma_{ui}} \right\}, \\
U_2(\varphi_u) &= \sum_{i=1}^{n} x_{2i}^\top \left\{ \frac{1}{2\sigma_{ui}^2} \left[\frac{\sigma_{wi}}{\sigma_{ui}} B_i - \frac{\sigma_{wi}^2}{\sigma_{ui}^2} - \frac{e_i}{\sigma_{ui}} - 1 \right] \right\} \sigma_{ui}^2, , \\
U_2(\varphi_w) &= \sum_{i=1}^{n} x_{3i}^\top \left\{ \frac{1}{2\sigma_{ui}^2} + \frac{1}{2\sigma_{wi}} \left[\frac{e_i}{\sigma_{wi}^2} - \frac{1}{\sigma_{ui}} \right] B_i \right\} \sigma_{wi}^2,
\end{aligned} \tag{4.10}$$

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

O gradiente da Eq. (4.7), ao assumir distribuição normal truncada para u_i, é dado por

$$
\begin{aligned}
U_1(\delta) &= -2\sum_{i=1}^{n} Z_i^\top \varepsilon_i, \\
U_2(\delta) &= -\sum_{i=1}^{n} Z_i^\top \left\{ \frac{e_i + \mu_i}{\sigma^2} + \frac{\lambda}{\sigma} D_i \right\} \eta^\top, \\
U_2(\eta) &= \sum_{i=1}^{n} \hat{\varepsilon}_i^\top \left\{ \frac{e_i + \mu_i}{\sigma^2} + \frac{\lambda}{\sigma} D_i \right\}, \\
U_2(\beta) &= \sum_{i=1}^{n} x_{1i}^\top \left\{ \frac{e_i + \mu_i}{\sigma^2} + \frac{\lambda}{\sigma} D_i \right\}, \\
U_2(\tau) &= \sum_{i=1}^{n} x_{2i}^\top \left\{ \frac{1}{\lambda\sigma} D_i - \frac{\sqrt{\lambda^{-2}+1}}{\sigma} C_i - \frac{e_i + \mu_i}{\sigma^2} \right\}, \\
U_2(\varphi_u) &= \sum_{i=1}^{n} \left\{ \frac{1}{2\sigma_u^2} \frac{\mu_i}{\sigma_u} C_i + \frac{1}{2\sigma^2} \left[\frac{(e_i + \mu_i)^2}{\sigma^2} - \frac{1}{\lambda\sigma} D_i \left(2\mu_i + e_i + \frac{\mu_i}{\lambda^2} \right) - 1 \right] \right\} \sigma_u^2, \\
U_2(\varphi_w) &= \sum_{i=1}^{n} \left\{ \frac{1}{2\sigma^2} \left[\frac{(e_i + \mu_i)^2}{\sigma^2} + \frac{\lambda}{\sigma} D_i \left(\mu_i + 2e_i + e_i\lambda^2 \right) - 1 \right] \right\} \sigma_w^2,
\end{aligned}
\tag{4.11}
$$

em que $C_i = \dfrac{\phi(c_i)}{\Phi(c_i)}$ com $c_i = \dfrac{\mu_i}{\sigma_u}$ e $D_i = \dfrac{\phi(d_i)}{\Phi(d_i)}$ com $d_i = \dfrac{\mu_i}{\sigma\lambda} - \dfrac{e_i\lambda}{\sigma}$.

Para obter esses gradientes assume-se função de ligação exponencial para σ_{ui}^2 e σ_{wi}^2 e que $\mu_i \in \mathbb{R}$.

4.6 Considerações finais

Este capítulo apresentou uma abordagem de máxima verossimilhança em dois estágios na presença de endogeneidade, assumindo distribuição seminormal, exponencial ou normal truncada para o termo de ineficiência e função de ligação exponencial para σ_{ui}^2 e σ_{wi}^2. Além disso, o gradiente analítico dos três modelos considerados foi apresentado.

O próximo capítulo trata da predição da eficiência técnica da unidade de produção na presença de endogeneidade.

Capítulo 5

Predição da eficiência técnica

O objetivo deste capítulo é apresentar uma forma de predição da eficiência técnica de cada produtor após obtenção das estimativas dos parâmetros do modelo por método de máxima verossimilhança em um ou dois estágios.

5.1 Considerações iniciais

Após obter estimativas dos parâmetros do modelo via método de máxima verossimilhança, o próximo passo é predizer a eficiência técnica de cada produtor. O termo de ineficiência, u_i, é uma variável aleatória de interesse particular neste cenário. As estimativas obtidas de $e_i = v_i - u_i$ contêm informação sobre u_i. Se $e_i > 0$, as chances são de que u_i não é grande, desde $\mathrm{E}(v_i) = 0$, o que sugere que esse produtor é relativamente eficiente, ao passo que se $e_i < 0$, é provável que u_i seja grande, o que sugere que esse produtor é relativamente ineficiente. O problema é extrair as informações que e_i contém de u_i. Uma solução é usar a distribuição condicional de u_i dado e_i, que possui qualquer informação que e_i contém relativa a u_i (Kumbhakar e Lovell, 2003).

5.2 Preditor proposto por Battese e Coelli (1995)

Na maioria das situações, o interesse está na eficiência do i-ésimo produtor, $\mathcal{E}_i = \exp(-u_i)$. Um preditor natural para essa quantidade é $\hat{\mathcal{E}}_i = \exp(-\hat{u}_i)$. No entanto, Battese e Coelli (1988) usaram $f(u_i|e_i)$ para derivar um preditor alternativo, que foi modificado por Battese e Coelli (1995) para levar em consideração a heterocedasticidade que pode existir em relação aos componentes de erro. Esse preditor alternativo é

$$\hat{\mathcal{E}}_i = E\{\exp(-u_i)|e_i\} = \left\{ \frac{\Phi(\mu_i^*/\sigma_i^* - \sigma_i^*)}{\Phi(\mu_i^*/\sigma_i^*)} \exp\left(\frac{1}{2}\sigma_i^{*2} - \mu_i^*\right) \right\}, \tag{5.1}$$

em que μ_i^* e σ_i^* variam conforme a especificação de u_i.

Para o modelo normal/seminormal μ_i^* e σ_i^* são

$$\begin{aligned} \mu_i^* &= \frac{-e_i\sigma_{ui}^2}{\sigma_i^2}, \\ \sigma_i^* &= \frac{\sigma_{wi}\sigma_{ui}}{\sigma_i}. \end{aligned} \tag{5.2}$$

Para o modelo normal/exponencial μ_i^* e σ_i^* são

$$\begin{aligned} \mu_i^* &= -e_i - \frac{\sigma_{wi}^2}{\sigma_{ui}}, \\ \sigma_i^* &= \sigma_{wi}. \end{aligned} \tag{5.3}$$

Para o modelo normal/normal truncada μ_i^* e σ^* são

$$\begin{aligned} \mu_i^* &= \frac{-e_i\sigma_u^2 + \mu_i\sigma_w^2}{\sigma^2}, \\ \sigma^* &= \frac{\sigma_w\sigma_u}{\sigma}. \end{aligned} \tag{5.4}$$

Battese e Coelli (1988) argumentam que, como a função de produção é geralmente definida pelo logaritmo da produção, $\ln y_i$, a eficiência técnica para a i-ésima unidade de produção deve ser definida como $E\{\exp(-u_i)|e_i\}$. Esse preditor pode ser considerado ótimo no sentido de minimizar o erro quadrático médio de predição.

Como os dados estão em termos de log, $\hat{\mathcal{E}}_i$ é uma medida da porcentagem pela qual uma unidade falha em alcançar a fronteira, que é taxa de produção ideal. Quanto mais próximo de um $\hat{\mathcal{E}}_i$ está, mais próximo o produtor está de atingir a produção ideal, com a tecnologia incorporada na função de produção.

5.3 Considerações finais

Este capítulo apresentou uma fórmula para a predição da eficiência técnica de cada produtor na presença de endogeneidade.

Testes de especificação e medidas de ajuste são apresentados no capítulo seguinte, com ênfase nos testes de endogeneidade.

Capítulo 6

Testes de especificação

O objetivo deste capítulo é apresentar alguns testes e medidas de ajuste usados na aplicação.

6.1 Considerações iniciais

Testes t, F, da razão de verossimilhança, de Wald e do multiplicador de Lagrange são apenas assintoticamente justificados na análise de fronteira estocástica. Portanto, estritamente falando, eles só podem ser usados quando o tamanho da amostra é grande. Infelizmente, os testes t e F não são mais justificados em amostras pequenas porque o erro composto no modelo de fronteira estocástica não é normalmente distribuído (Coelli et al., 2005).

Além de testar hipóteses sobre $\boldsymbol{\beta}$, os pesquisadores de fronteira estocástica estão frequentemente interessados em testar a ausência dos efeitos de ineficiência. No caso dos modelos seminormal e exponencial, a hipótese nula é uma única restrição envolvendo um único parâmetro. Se os parâmetros do modelo foram estimados pelo método de máxima verossimilhança, é possível testar tal hipótese usando um teste z simples (porque os estimadores de máxima verossimilhança não restritos são assintoticamente normalmente distribuídos). Embora seja possível testar os efeitos de ineficiência usando um teste de Wald, um teste do multiplicador de Lagrange ou um teste da razão de verossimilhança (TRV), a natureza unilateral da hipótese alternativa implica que esses testes são difíceis de interpretar. Além disso, eles não têm distribuições assintóticas qui-quadrado (Coelli et al., 2005).

Procedimentos de teste do tipo descrito acima também podem ser usados para testar hipóteses sobre os parâmetros de outros modelos de fronteira. Por exemplo, no caso do modelo normal truncada, a hipótese nula: $H_0 : \mu = \sigma_u^2 = 0$ pode ser testada. Além de poder usar estimativas do modelo normal truncada para testar a hipótese nula de que o modelo mais simples, seminormal, é adequado. As hipóteses nula e alternativa relevantes são $H_0 : \mu = 0$ e $H_1 : \mu \neq 0$. Novamente, se os parâmetros do modelo foram estimados pelo método de máxima verossimilhança, é possível usar o teste z ou o teste da razão de verossimilhança (Coelli et al., 2005).

6.2 Testes de endogeneidade

É possível testar a significância conjunta dos componentes do termo η nos modelos de um e dois estágios. Se os componentes forem conjuntamente significativos, conclui-se que há endogeneidade no modelo. Quando os componentes não são conjuntamente significativos, isso indicaria que o termo de correção não é necessário e a eficiência pode ser estimada pelos modelos de fronteira convencionais. Desta forma, se a endogeneidade estiver presente, pode ser benéfico usar a abordagem de variáveis instrumentais, pois o viés pode ser menor. Sendo que, as magnitudes relativas dos vieses, ao usar o modelo endógeno ou exógeno, dependem do grau de endogeneidade e do problema de identificação (Karakaplan e Kutlu, 2015).

Para tanto, alguns testes podem ser usados para testar a presença de endogeneidade, tais como um teste de Wald ou um TRV. O teste de Wald requer apenas o estimador irrestrito, enquanto que o TRV requer o cálculo de estimadores restritos e irrestritos.

6.2.1 Teste de Wald

Sob H_0, o teste de Wald é dado por:

$$\begin{bmatrix} \hat{\eta}_1 & \cdots & \hat{\eta}_p \end{bmatrix} \begin{bmatrix} Var\left(\hat{\eta}_1\right) & \cdots & Cov\left(\hat{\eta}_1, \hat{\eta}_p\right) \\ \vdots & \ddots & \vdots \\ Cov\left(\hat{\eta}_p, \hat{\eta}_1\right) & \cdots & Var\left(\hat{\eta}_p\right) \end{bmatrix}^{-1} \begin{bmatrix} \hat{\eta}_1 \\ \vdots \\ \hat{\eta}_p \end{bmatrix} \sim \chi^2_{(p)}, \tag{6.1}$$

no qual p representa o número de variáveis presumidas endógenas. Rejeita-se H_0 quando o p-valor desse teste $\chi^2_{(p)}$ for menor que o nível de significância considerado.

6.2.2 Teste da razão de verossimilhança

Sob H_0, o TRV é da forma:

$$-2(\ln \hat{L}_{\hat{\eta}_i=0} - \ln \hat{L}_{\hat{\eta}_i\neq 0}) \sim \chi^2_{(p)}, \quad i = 1, \ldots, p, \tag{6.2}$$

no qual p representa o número de variáveis presumidas endógenas. Rejeita-se H_0 quando o p-valor desse teste $\chi^2_{(p)}$ for menor que o nível de significância considerado.

Em alguns problemas, um destes estimadores pode ser muito mais fácil de calcular do que o outro. Como consequência, a escolha entre eles é tipicamente feita com base na facilidade computacional.

Note que quando $\eta = 0$, os erros padrão do segundo estágio do estimador de dois estágios são válidos e, assintoticamente, eles são tão eficientes quanto a versão de um estágio.

6.3 Validade dos instrumentos

Ainda é possível testar se os instrumentos são exógenos e relevantes, ou seja, se os instrumentos não possuem correlação com o erro (condição de exogeneidade) e se são altamente correlacionados com os regressores endógenos (condição de relevância). Idealmente, os instrumentos não apenas se correlacionam com os regressores endógenos, mas também podem explicar uma grande parte da variação deles, indicando que eles não são instrumentos fracos.

Se os instrumentos são endógenos, a inconsistência no estimador de variáveis instrumentais se torna arbitrariamente grande à medida que a correlação entre o instrumento e a variável endógena se aproxima de zero. O que faz com que as estimativas obtidas com o uso de variáveis instrumentais não sejam mais consistentes. Assim, correlações aparentemente pequenas entre os instrumentos e o termo de erro podem causar inconsistência severa - e, portanto, severo viés de amostra finita - se a variável instrumental é apenas fracamente correlacionada com a variável

endógena (Wooldridge, 2010).

Para verificar se os instrumentos usados são válidos, pode-se calcular as correlações entre eles e as variáveis para quais estão servindo de instrumentos, bem como, calcular as correlações das variáveis instrumentais com o erro do modelo. Se as condições de exogeneidade e relevância são atendidas, os instrumentos são ditos válidos.

6.4 Medidas de ajuste

Para fins de comparação e ajuste de modelos é útil se obter as seguintes medidas (Greene, 2012a): critério de informação de Akaike (AIC) e critério de informação Bayesiano (BIC), bem como, correlação de Pearson, viés e raiz do erro quadrático médio (REQM) entre os valores observados e estimados da variável resposta.

6.5 Considerações finais

Este capítulo apresentou alguns testes de hipóteses de parâmetros e da presença de endogeneidade para a análise de fronteira estocástica. Além de formas de verificar à validade dos instrumentos usados e medidas para comparação e ajuste de modelos.

No próximo capítulo são apresentados os resultados das técnicas e modelos discutidos a um conjunto de dados reais.

Capítulo 7

Aplicação

O objetivo deste capítulo é apresentar e discutir os resultados das técnicas e modelos ilustrados nos capítulos anteriores, para quando presume-se a existência de variáveis endógenas, a um conjunto de dados reais.

7.1 Introdução

Seguindo as técnicas descritas nos capítulos anteriores foi realizado o ajuste de seis modelos. Os parâmetros dos modelos considerados são estimados em um (Cap. 3) ou dois estágios (Cap. 4), presumindo-se distribuição seminormal, exponencial ou normal truncada para o termo de ineficiência, u_i. O processo de modelagem postula uma representação Cobb-Douglas, em uma abordagem típica de fronteira estocástica e é realizado sob a hipótese de endogeneidade das variáveis assistência técnica (*assistec*) e acesso a crédito (*finan*), como assumido em Souza, Gomes e Alves (2018). Estas são variáveis complexas que podem envolver muitos fatores relativos a estrutura da unidade de produção e são fortes candidatas a endogeneidade.

Os modelos seguem as linhas básicas de Karakaplan e Kutlu (2015), em que os parâmetros são estimados em um único estágio (abordagem MVIC). Além de modelos em que os parâmetros são estimados em dois estágios (abordagem MVIL), usando a correção da variância proposta por Murphy e Topel (1985).

Como a linguagem *R* ainda não possui pacote disponível para se obter estimativas consis-

tentes dos parâmetros da análise de fronteira estocástica na presença de variáveis endógenas, de forma mais geral, realizou-se a criação de comandos para permitir o uso de modelos na presença ou não de endogeneidade e a heterocedasticidade em relação a ambos os termos de erro, considerando as três distribuições mais usuais para u_i e que v_i tem distribuição normal.

Ao se especificar as distribuições normal/seminormal ou normal/exponencial é possível o ajuste de modelos com componentes de erro heterocedásticos, condicional a um conjunto de covariáveis, em que $\sigma^2_{ui} = \exp(x^\top_{2i}\varphi_u)$ e $\sigma^2_{wi} = \exp(x^\top_{3i}\varphi_w)$. Enquanto que para uma distribuição normal/normal truncada é possível ajustar um modelo de média condicional, sendo a média modelada como uma função linear de um conjunto de covariáveis, $\mu_i = x^\top_{2i}\tau$.

7.2 Descrição das variáveis

O modelo de produção assume como variável dependente a renda bruta dos estabelecimentos rurais dos municípios brasileiros em reais (*y*), isto é, o valor total da produção agropecuária dos estabelecimentos e, como insumos, as despesas com terra (*terra*), trabalho (*trab*) e capital (*cap*, insumos tecnológicos). Estas variáveis foram extraídas da base de dados do censo agropecuário brasileiro do IBGE de 2006 e foram agregadas em nível municipal. Maiores informações sobre as variáveis em Souza, Gomes e Alves (2016).

Como variáveis contextuais que afetam a produção consideram-se indicadores agregados referentes as características sociais (*social*), demográficas (*demo*) e ambientais (*ambi*) do desenvolvimento rural dos municípios. Além de variáveis referentes ao acesso a crédito (*finan*, proporção de financiamento obtido pelos estabelecimentos rurais no município), a assistência técnica (*assistec*, proporção de agricultores que receberam assistência técnica no município), um indicador de concentração de renda por município (*gini*) e uma *dummy* regional (*regiao*, uma variável indicando a região do município). As variáveis contextuais foram extraídas da base de dados do censo demográfico brasileiro de 2010, do Ministério da Saúde, data base 2011, e do INEP, data base 2009.

Com exceção de região, as demais variáveis contextuais variam no intervalo de zero a um, com um indicando valores mais altos. Essas variáveis contextuais foram ordenadas e norma-

lizadas pelo máximo. Essa abordagem empresta propriedades estatísticas não paramétricas à análise e contorna problemas associados à presença de atipicidades (*outliers*) e à escala de operação (heteroscedasticidade) (Souza, Gomes e Alves, 2016).

Para a função de produção considera-se o logaritmo da variável *y*, como variável resposta, e como variáveis explicativas, o logaritmo dos fatores de produção - *terra*, *trab* e *cap* - e a *dummy* regional. As variáveis *assistec*, *finan* e *gini* são usadas para modelar a função de σ_{ui}^2 para o nível de ineficiência técnica dos estabelecimentos agropecuários. As variáveis *social*, *demo* e *ambi* são variáveis instrumentais externas na análise. Assume-se que as variáveis *assistec* e *finan* são potencialmente endógenas.

Estatísticas descritivas dessas variáveis são apresentadas na Tabela 7.1. A Figura 7.1 ilustra o *box plot* de cada uma das variáveis, com as variáveis *y*, *terra*, *trab* e *cap* na forma logarítmica.

Tabela 7.1: Estatísticas descritivas das variáveis.

Variável	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo	CV (%)
renda	696	10191	25983	71974	59968	5970271	315.13
terra	21	1435	5207	15269	14734	2906109	376.22
trab	165	3176	7920	34901	17711	9538791	730.41
cap	220	5424	15231	48230	35312	12963405	626.16
finan	0.01	0.25	0.50	0.50	0.75	1.00	57.73
social	0.02	0.27	0.52	0.51	0.74	0.98	49.83
demo	0.17	0.39	0.49	0.49	0.59	0.95	26.24
ambi	0.17	0.45	0.52	0.51	0.59	0.86	18.53
assistec	0.001	0.25	0.50	0.50	0.76	1.00	58.24
regiao	1	2	3	2.82	4	5	36.11
gini	0.53	0.73	0.80	0.79	0.86	1.00	12.20

A renda bruta média da produção agropecuária agregada no nível municipal é de R$ 71974 com um coeficiente de variação de 315%, o que indica a grande variabilidade de renda entre os estabelecimentos agropecuários dos municípios. A Tabela 7.1 mostra que a despesa com capital é o componente mais dominante dos custos, seguidos por trabalho e terra, respectivamente. O índice de concentração de renda municipal (*gini*) em média é alto (79%), indicando uma alta variabilidade entre as rendas brutas rurais municipais. A média municipal de assistência técnica (*assistec*) e acesso a crédito agrícola (*finan*) é de 50%, sendo que alguns municípios pratica-

mente não tiveram acesso algum a financiamentos e assistências técnicas rurais. Os indicadores agregados referentes às características sociais, demográficas e ambientais do desenvolvimento rural dos municípios apresentaram valores médios em torno de 50%, com coeficientes de variação de aproximadamente 50%, 26% e 19%, respectivamente.

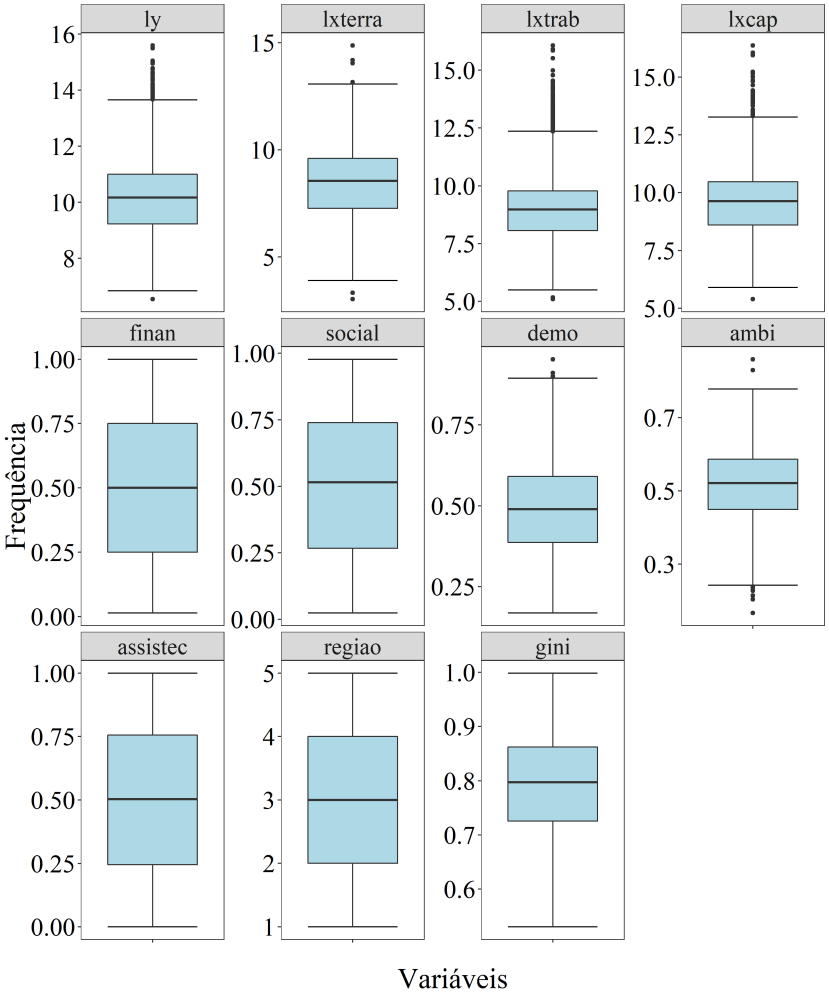

Figura 7.1: *Box plot* das variáveis.

7.3 Processo de escolha dos modelos

A Tabela 7.2 mostra algumas medidas de ajuste para a escolha dos modelos de um e dois estágios. Especificamente, apresenta os valores de log-verossimilhança ($\ln L$), critério de informação de Akaike (AIC) e critério de informação bayesiano (BIC) dos modelos. Além dos

valores de correlação de Pearson, viés e raiz do erro quadrático médio (REQM) entre os valores observados e estimados da variável resposta, $\ln(renda)$.

Ao se considerar a distribuição exponencial para u_i não houve convergência pelo método de otimização BFGS[1], que usa os gradientes analíticos apresentados nas seções 3.4 e 4.5[2]. Para tanto, foi utilizado o método de Nelder-Mead, que não usa os gradientes analíticos e, portanto, é mais lento. Foram necessárias 53280 iterações até a convergência das estimativas do modelo normal/exponencial usando a abordagem MVIC, enquanto que a abordagem MVIL, apenas 4514. Em ambos os casos, em comparação as outras duas distribuições, a velocidade de convergência ao assumir uma distribuição exponencial para u_i é consideravelmente mais lenta.

Tabela 7.2: Medidas de ajuste.

Abordagem	Distribuição de $\boldsymbol{u}$	$\ln L$	AIC	BIC	$\mathrm{Cor}(\boldsymbol{y}, \hat{\boldsymbol{y}})$	Viés	REQM
MVIC	Seminormal	-874.8	1827.5	2081.4	0.8792	0.19	1.135
	Exponencial	-1330.6	2739.3	2993.2	0.8769	0.32	1.138
	Normal truncada	-432.2	944.5	1204.9	0.7315	2.87	2.895
MVIL	Seminormal	-892.1	1814.3	1911.9	0.8792	0.20	1.126
	Exponencial	-1284.7	2599.3	2697.0	0.8772	0.26	1.199
	Normal truncada	-435.7	903.5	1007.6	0.7636	2.88	2.901

Embora não tenha ocorrido problemas de convergência ao usar o método de otimização de Nelder-Mead, para os modelos normal/exponencial, e do BFGS, para os modelos normal/seminormal e normal/normal truncada, na obtenção das estimativas dos parâmetros dos seis modelos considerados (Tabela 7.2), ressalta-se que pode haver problemas de convergência dependendo da própria distribuição assumida para u_i e do número de variáveis incluídas nos efeitos de ineficiência.

Quanto às medidas de ajuste, os modelos normal/normal truncada obtidos via abordagens MVIC e MVIL apresentaram os menores valores de AIC e BIC e maiores valores de log-verossimilhança, sendo os modelos a serem escolhidos por esses critérios. Contudo, os mo-

[1] O método BFGS é um método quase-Newton, especificamente aquele publicado simultaneamente em 1970 por Broyden, Fletcher, Goldfarb e Shanno. Esse método usa valores de função e gradientes para criar uma imagem da superfície a ser otimizada.

[2] Para obtenção dos erros padrão, a matriz hessiana foi obtida numericamente.

delos normal/seminormal ajustaram-se melhor ao se considerar os critérios de maior correlação de Pearson e menor valor de viés e REQM entre os valores observados e estimados da variável resposta, $\ln(renda)$. As correlações de Pearson entre os valores observados e estimados das variáveis presumidas endógenas (*assistec* e *finan*) são todas maiores que 0.8, nos seis modelos considerados, indicando o bom ajuste destas variáveis estimadas via regressões lineares por variáveis instrumentais.

Assim, com base nos resultados da Tabela 7.2, referentes à qualidade de ajuste, e pelo princípio da parcimônia, foram escolhidos os modelos normal/seminormal, com seus respectivos parâmetros estimados em um ou dois estágios para modelar a renda bruta rural total dos municípios brasileiros.[3]

O modelo de fronteira estocástica de produção normal/seminormal ajustado é da forma

$$
\begin{aligned}
\ln(y_i) &= \beta_0 + \beta_1 \ln(terra_i) + \beta_2 \ln(trab_i) + \beta_3 \ln(cap_i) + \beta_4 regiao_{norte_i} + \beta_5 regiao_{nordeste_i} + \\
&\quad + \beta_6 regiao_{sudeste_i} + \beta_7 regiao_{sul_i} + v_i - u_i, \\
\ln(\sigma^2_{ui}) &= \varphi_{u0} + \varphi_{u1} assistec_i + \varphi_{u2} finan_i + \varphi_{u3} gini_i, \\
\ln(\sigma^2_{wi}) &= \varphi_{w0}, \\
assistec_i &= \delta_0 + \delta_1 \ln(terra_i) + \delta_2 \ln(trab_i) + \delta_3 \ln(cap_i) + \delta_4 regiao_{norte_i} + \delta_5 regiao_{nordeste_i} + \\
&\quad + \delta_6 regiao_{sudeste_i} + \delta_7 regiao_{sul_i} + \delta_8 social_i + \delta_9 demo_i + \delta_{10} ambi_i + \delta_{11} gini_i + \varepsilon_{1i}, \\
finan_i &= \gamma_0 + \gamma_1 \ln(terra_i) + \gamma_2 \ln(trab_i) + \gamma_3 \ln(cap_i) + \gamma_4 regiao_{norte_i} + \gamma_5 regiao_{nordeste_i} + \\
&\quad + \gamma_6 regiao_{sudeste_i} + \gamma_7 regiao_{sul_i} + \gamma_8 social_i + \gamma_9 demo_i + \gamma_{10} ambi_i + \gamma_{11} gini_i + \varepsilon_{2i}.
\end{aligned}
$$

Para comparar as regiões, a região Centro-Oeste foi definida como o nível base.

7.4 Resultados

As Tabelas 7.3 e 7.4 mostram as estimativas para o modelo normal/seminormal de informação completa. A Tabela 7.5 apresenta os resultados obtidos no primeiro estágio do modelo de

[3] Consulte os Apêndices A e B para verificar os resultados obtidos ao se especificar as distribuições exponencial e normal truncada para u_i.

Tabela 7.3: Regressão linear por variáveis instrumentais da assistência técnica e acesso a crédito estimadas em um estágio assumindo que $u_i \sim \mathcal{N}^+(0, \sigma^2_{ui})$.

Variável	Coeficiente	EP	z	P-valor	LI	LS
assistec						
constante	-0.260	0.036	-7.277	0.000	-0.331	-0.190
$\ln(terra)$	0.005	0.004	1.247	0.212	-0.003	0.013
$\ln(trab)$	0.010	0.003	3.071	0.002	0.004	0.017
$\ln(cap)$	0.082	0.005	17.294	0.000	0.073	0.091
regiao$_{norte}$	0.048	0.015	3.129	0.002	0.018	0.078
regiao$_{nordeste}$	0.048	0.015	3.201	0.001	0.019	0.078
regiao$_{sudeste}$	0.071	0.013	5.292	0.000	0.045	0.097
regiao$_{sul}$	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
demo	-0.016	0.028	-0.560	0.575	-0.071	0.040
ambi	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
finan						
constante	-0.508	0.036	-13.939	0.000	-0.580	-0.437
$\ln(terra)$	0.027	0.004	6.831	0.000	0.019	0.035
$\ln(trab)$	-0.005	0.003	-1.381	0.167	-0.011	0.002
$\ln(cap)$	0.129	0.005	26.873	0.000	0.120	0.138
regiao$_{norte}$	-0.076	0.015	-4.884	0.000	-0.106	-0.045
regiao$_{nordeste}$	-0.080	0.015	-5.241	0.000	-0.110	-0.050
regiao$_{sudeste}$	-0.057	0.014	-4.223	0.000	-0.084	-0.031
regiao$_{sul}$	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
demo	-0.222	0.029	-7.576	0.000	-0.279	-0.165
ambi	-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

dois estágios das regressões lineares por variáveis instrumentais para as variáveis presumidas endógenas. Na Tabela 7.6 estão as estimativas obtidas no segundo estágio do modelo de dois estágios, com a correção da variância de Murphy e Topel (1985).

Nas Tabelas 7.3 e 7.5 note a significância (p-valor < 0.05) das variáveis *social* e *gini*. Em ambos os casos, as condições sociais afetam positivamente, enquanto que a concentração de renda tem um efeito negativo sobre as variáveis assistência técnica e acesso a crédito. Assim, uma melhoria do indicador social e uma menor concentração de renda tenderão a facilitar o acesso a assistência técnica e a crédito rural. Já as condições demográficas e ambientais têm um

Tabela 7.4: Parâmetros do modelo normal/seminormal estimados em um estágio.

Variável	Coeficiente	EP	z	P-valor	LI	LS
Fronteira						
constante	2.477	0.125	19.852	0.000	2.233	2.722
$\ln(terra)$	0.110	0.014	7.639	0.000	0.082	0.139
$\ln(trab)$	0.206	0.011	17.945	0.000	0.184	0.229
$\ln(cap)$	0.509	0.018	28.977	0.000	0.475	0.544
regiao$_{norte}$	0.066	0.055	1.201	0.230	-0.042	0.174
regiao$_{nordeste}$	0.129	0.053	2.425	0.015	0.025	0.233
regiao$_{sudeste}$	0.242	0.046	5.211	0.000	0.151	0.333
regiao$_{sul}$	0.335	0.048	7.003	0.000	0.241	0.429
$\ln(\boldsymbol{\sigma_u^2})$						
constante	5.637	0.201	28.040	0.000	5.243	6.031
assistec	0.931	0.525	1.772	0.076	-0.099	1.961
finan	-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
$\ln(\boldsymbol{\sigma_w^2})$						
constante	-1.078	0.023	-47.806	0.000	-1.122	-1.034
$\boldsymbol{\eta_{assistec}}$						
constante	0.890	0.079	11.194	0.000	0.734	1.045
$\boldsymbol{\eta_{finan}}$						
constante	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

efeito negativo significativo apenas sobre o acesso a crédito, o que pode acontecer devido, em geral, ser necessário um maior gasto para produzir preservando o meio-ambiente.

As estimativas dos gastos com terra, trabalho e capital, nas Tabelas 7.4 e 7.6, têm efeitos positivos e significativos sobre a renda, bem como, com exceção da região Norte, as demais regiões têm efeitos positivos significativos sobre a renda, quando comparadas à região Centro-Oeste.

Note ainda que os componentes de acesso a crédito (*finan*) e de concentração de renda (*gini*), presentes nas Tabelas 7.4 e 7.6, têm efeitos negativos e significativos na função de σ_{ui}^2 para o nível de ineficiência técnica dos estabelecimentos agropecuários, bem como o fornecimento de assistência técnica (*assistec*) não tem um efeito significativo a 5%, isto é, um maior acesso a crédito rural e a concentração de renda diminuem a ineficiência dos estabelecimentos agrope-

Tabela 7.5: Regressão linear por variáveis instrumentais da assistência técnica e acesso a crédito estimadas em dois estágios.

Variável	Coeficiente	EP	t	P-valor	LI	LS
assistec						
constante	-0.293	0.036	-8.148	0.000	-0.363	-0.222
$\ln(terra)$	0.003	0.004	0.864	0.388	-0.004	0.011
$\ln(trab)$	0.004	0.003	1.167	0.243	-0.003	0.010
$\ln(cap)$	0.079	0.005	16.716	0.000	0.070	0.088
regiao$_{norte}$	0.055	0.015	3.598	0.000	0.025	0.085
regiao$_{nordeste}$	0.047	0.015	3.124	0.002	0.018	0.077
regiao$_{sudeste}$	0.059	0.013	4.419	0.000	0.033	0.085
regiao$_{sul}$	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
demo	-0.003	0.029	-0.109	0.913	-0.060	0.054
ambi	-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
finan						
constante	-0.521	0.036	-14.313	0.000	-0.593	-0.450
$\ln(terra)$	0.027	0.004	6.682	0.000	0.019	0.035
$\ln(trab)$	-0.007	0.003	-2.186	0.029	-0.014	-0.001
$\ln(cap)$	0.128	0.005	26.669	0.000	0.118	0.137
regiao$_{norte}$	-0.073	0.015	-4.698	0.000	-0.103	-0.042
regiao$_{nordeste}$	-0.081	0.015	-5.269	0.000	-0.111	-0.051
regiao$_{sudeste}$	-0.062	0.014	-4.588	0.000	-0.089	-0.036
regiao$_{sul}$	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
demo	-0.217	0.029	-7.359	0.000	-0.275	-0.159
ambi	-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

cuários, enquanto que o fornecimento de assistência técnica não influencia significativamente o nível de ineficiência. Observe que há evidência de endogeneidade em ambos os casos ($\hat{\eta}_{finan}$ e $\hat{\eta}_{assistec}$ com p-valor < 0.05).

Os resultados dos testes de Wald (Eq. 6.1) e da razão de verossimilhança (Eq. 6.2) sobre a presença de endogeneidade são apresentados na Tabela 7.7. Note que em ambas as abordagens é rejeitada a hipótese nula de exogeneidade. Logo, há evidências da presença de endogeneidade. Consequentemente, para se obter estimativas consistentes dos parâmetros, faz sentido trabalhar com modelos que levam em conta a endogeneidade.

Tabela 7.6: Parâmetros do modelo normal/seminormal estimados em dois estágios.

Variável	Coeficiente	EP	z	P-valor	LI	LS
Fronteira						
constante	2.552	0.126	20.306	0.000	2.306	2.799
$\ln(terra)$	0.111	0.015	7.521	0.000	0.082	0.140
$\ln(trab)$	0.202	0.012	16.813	0.000	0.178	0.226
$\ln(cap)$	0.507	0.018	28.133	0.000	0.472	0.543
regiao$_{norte}$	0.062	0.055	1.123	0.261	-0.046	0.170
regiao$_{nordeste}$	0.131	0.053	2.455	0.014	0.026	0.235
regiao$_{sudeste}$	0.231	0.046	4.995	0.000	0.140	0.321
regiao$_{sul}$	0.319	0.047	6.736	0.000	0.226	0.412
$\ln(\boldsymbol{\sigma_u^2})$						
constante	6.337	0.594	10.662	0.000	5.172	7.502
assistec	-0.182	0.578	-0.315	0.753	-1.314	0.950
finan	-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
$\ln(\boldsymbol{\sigma_w^2})$						
constante	-1.065	0.023	-45.429	0.000	-1.111	-1.019
$\boldsymbol{\eta_{assistec}}$						
constante	0.664	0.080	8.335	0.000	0.508	0.820
$\boldsymbol{\eta_{finan}}$						
constante	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

Tabela 7.7: Teste de Wald e TRV para a presença de endogeneidade.

Abordagem	Wald	TRV	P-valor
MVIC	201	342	0.000
MVIL	149	208	0.000

A Figura 7.2 ilustra os *box plots* por região para as classificações normalizadas das medições de eficiência técnica preditas, $\hat{\mathcal{E}}_i$, pelos modelos normal/seminormal via abordagens MVIC e MVIL. Note que há uma dominação da região Centro-Oeste sobre as demais regiões e que o Nordeste apresenta os menores níveis de eficiência técnica.

A Tabela 7.8 resume a importância relativa dos fatores de produção, incluindo os retornos à escala, para os modelos de um e dois estágios. Em ambos os casos, o capital (insumos tecnológicos) domina, seguido de trabalho e terra, mostrando que o capital como insumo tem maior

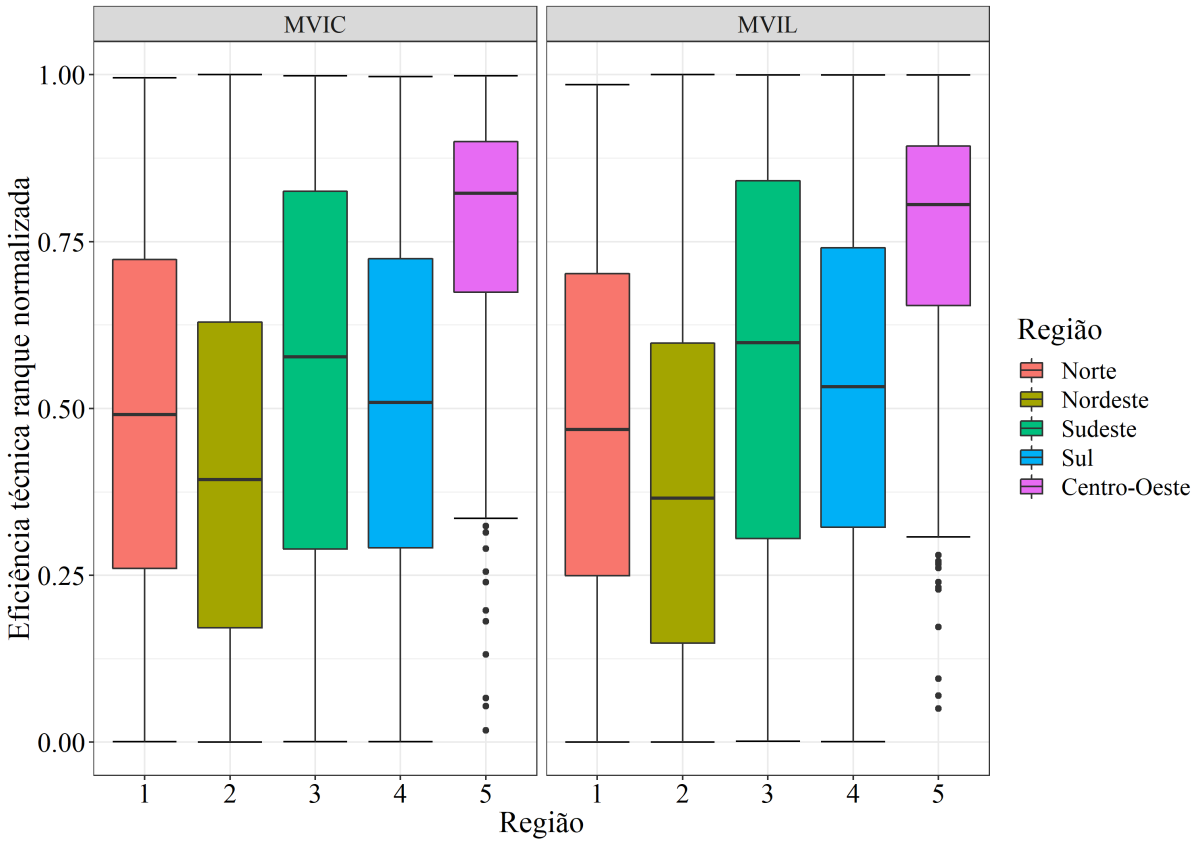

Figura 7.2: *Box plot* normalizado das eficiências técnicas preditas pelos modelos normal/semi-normal via abordagens MVIC e MVIL.

influência na produção, o que concorda com a literatura existente.

$$\begin{aligned} H_0 &: \beta_1 + \beta_2 + \beta_3 = 1 \\ H_1 &: \beta_1 + \beta_2 + \beta_3 < 1 \\ t &= \frac{\beta_1 + \beta_2 + \beta_3 - 1}{\text{EP}(\beta_1 + \beta_2 + \beta_3)} \sim t_{(n)}. \end{aligned} \tag{7.1}$$

Conforme o teste t unilateral da Eq. (7.1), em ambas as abordagens, a tecnologia mostra retornos decrescentes à escala, no qual as estatísticas de teste são iguais a $t = -16.84$ e $t = -17.90$ e os p-valores menores que 0.01%. Consequentemente, os retornos à escala de 0.83 e 0.82, respectivamente, indicam retornos decrescentes à escala.

A Tabela 7.9 apresenta as correlações referentes à validade das variáveis instrumentais usa-

Tabela 7.8: Elasticidades relativas e retornos à escala.

Fator de produção	Coeficiente	EP	t	P-valor	LI	LS
MVIC						
Terra	0.13	0.02	7.69	0.00	0.10	0.17
Trabalho	0.25	0.01	18.30	0.00	0.22	0.28
Capital	0.62	0.02	31.01	0.00	0.58	0.66
Retornos à escala	0.83	0.01				
MVIL						
Terra	0.14	0.02	7.91	0.00	0.10	0.17
Trabalho	0.25	0.01	18.13	0.00	0.22	0.27
Capital	0.62	0.02	31.47	0.00	0.58	0.66
Retornos à escala	0.82	0.01				

das. Observe que, com exceção da variável *gini*, os instrumentos possuem correlação de moderada à alta com as variáveis que estão instrumentando (*finan* e *assistec*), além de possuírem baixas correlações com os erros dos dois modelos. Logo, o conjunto geral de instrumentos são bons, atendendo as condições de exogeneidade e relevância, sendo portanto, válidos.

Tabela 7.9: Correlações de Pearson entre os instrumentos e os regressores endógenos e entre os instrumentos e os erros dos modelos, obtidas pelas abordagens MVIC e MVIL.

Variável	$\mathrm{Cor}(\boldsymbol{Z}, \boldsymbol{x}_{\text{assistec}})$	$\mathrm{Cor}(\boldsymbol{Z}, \boldsymbol{x}_{\text{finan}})$	$\mathrm{Cor}(\boldsymbol{Z}, \hat{\boldsymbol{e}}_{\boldsymbol{MVIC}})$	$\mathrm{Cor}(\boldsymbol{Z}, \hat{\boldsymbol{e}}_{\boldsymbol{MVIL}})$
$\ln(terra)$	0.65	0.71	0.08	0.10
$\ln(trab)$	0.52	0.56	0.09	0.11
$\ln(cap)$	0.67	0.75	0.09	0.10
regiao_{norte}	-0.11	-0.07	-0.003	-0.005
$\text{regiao}_{nordeste}$	-0.63	-0.57	-0.04	-0.06
$\text{regiao}_{sudeste}$	0.26	0.17	0.02	0.03
regiao_{sul}	0.46	0.40	0.01	0.02
social	0.77	0.63	0.14	0.15
demo	0.54	0.48	0.13	0.15
ambi	0.51	0.38	0.01	0.02
gini	-0.09	0.04	0.37	0.37

7.5 Conclusão

Ao se considerar uma especificação normal/seminormal para os termos de erro do modelo de fronteira estocástica de produção, as estimativas obtidas dos modelos para lidar com a endoge-

neidade pelo método de máxima verossimilhança em um estágio (abordagem MVIC), em geral, concordam com as obtidas pelo método de máxima verossimilhança em dois estágios (abordagem MVIL). Há evidências de endogeneidade do acesso a crédito e da assistência técnica em ambos os casos. Veja as Tabelas 7.4, 7.6 e 7.7. As eficiências técnicas preditas pelas duas abordagens também são semelhantes (Figura 7.2).

Como os resultados das duas abordagens são semelhantes recomenda-se o uso dos modelos de um estágio, uma vez que, sob as condições de regularidade padrões dos estimadores de máxima verossimilhança, o estimador MVIC é mais eficiente que o estimador MVIL, no sentido de, no geral, produzir os menores desvios padrão. Assim, para casos computacionalmente intensivos, em que não haja convergência pelo modelo de um estágio, utiliza-se o modelo de dois estágios.

Em um trabalho futuro, devido a correlação da variável *gini* com o erro (Tabela 7.9), pode ser interessante incluir a variável indicadora de concentração de renda como uma possível variável endógena ou não utilizá-la como uma variável instrumental.

Capítulo 8

Considerações finais

Com base na literatura disponível e na linguagem *R* - a qual não possui pacote ainda disponível para tratar da presença de endogeneidade na análise de fronteira estocástica, de forma a permitir covariáveis nos componentes de ruído e ineficiência - tornou-se possível a criação de comandos para a análise de fronteira estocástica via método de máxima verossimilhança, com o uso de três diferentes parametrizações para o termo de ineficiência, regressores exógenos ou endógenos e heteroscedasticidade em relação aos termos de erro das especificações seminormal e exponencial e na média da especificação normal truncada.

Outra contribuição deste trabalho é a expressão das fórmulas dos gradientes analíticos dos modelos de um e dois estágios, considerando uma regressão linear por variáveis instrumentais para as variáveis presumidas endógenas.

Além disso, implementou-se a predição da eficiência técnica e testes de endogeneidade, sendo a estimação dos parâmetros via método de máxima verossimilhança em um estágio, baseado em Karakaplan e Kutlu (2015) e método de máxima verossimilhança em dois estágios, com correção da variância de Murphy e Topel (1985), através do uso de variáveis instrumentais e gradientes analíticos, com a estimação em dois estágios apresentando menores problemas para convergência.

As conclusões da aplicação de uma fronteira estocástica de produção normal/seminormal sob endogeneidade aos dados municipais do censo agropecuário brasileiro de 2006 são nota-

velmente similares. Havendo convergência nos modelos de um e dois estágios, recomenda-se o uso da abordagem de máxima verossimilhança em um estágio, por ser mais eficiente.

Salienta-se que a correção da matriz de variância-covariância de Murphy e Topel (1985) pode alterar a significância das estimativas de variáveis importantes em relação às estimativas de um estágio, assim como pode alterar as eficiências técnicas preditas. Isso pode ser observado com maior intensidade ao se considerar um modelo normal/exponencial, o qual costuma apresentar problemas de convergência.

Mais estudos precisam ser feitos para a criação de comandos que possibilitem outras parametrizações para o termo de ineficiência, tal como a gama, ou ainda o uso de regressões não-lineares por variáveis instrumentais para as variáveis presumidas endógenas, além de diferentes análises diagnósticas, incluindo análise de resíduos. Contudo, no momento, almeja-se a disponibilização de uma rotina e/ou pacote no *R* para as abordagens e testes descritos neste trabalho.

Bibliografia

Aigner, Dennis, Lovell, CA Knox e Schmidt, Peter (1977). "Formulation and estimation of stochastic frontier production function models". Em: *Journal of econometrics* 6.1, pp. 21–37.

Amsler, Christine, Prokhorov, Artem e Schmidt, Peter (2016). "Endogeneity in stochastic frontier models". Em: *Journal of Econometrics* 190.2, pp. 280–288.

Andrade, Bernardo B de e Souza, Geraldo S (2017). "Likelihood computation in the normal-gamma stochastic frontier model". Em: *Computational Statistics*, pp. 1–16.

Battese, George E e Coelli, Tim J (1988). "Prediction of firm-level technical efficiencies with a generalized frontier production function and panel data". Em: *Journal of econometrics* 38.3, pp. 387–399.

— (1992). "Frontier production functions, technical efficiency and panel data: with application to paddy farmers in India". Em: *Journal of productivity analysis* 3.1-2, pp. 153–169.

Battese, George Edward e Coelli, Tim J (1995). "A model for technical inefficiency effects in a stochastic frontier production function for panel data". Em: *Empirical economics* 20.2, pp. 325–332.

Belotti, Federico et al. (2013). "Stochastic frontier analysis using Stata". Em: *The Stata Journal* 13.4, pp. 719–758.

Broyden, Charles George (1970). "The convergence of a class of double-rank minimization algorithms 1. general considerations". Em: *IMA Journal of Applied Mathematics* 6.1, pp. 76–90.

Coelli, Timothy J et al. (2005). *An introduction to efficiency and productivity analysis*. Springer Science & Business Media.

Davidson, Russell e MacKinnon, James G (2004). *Econometric theory and methods*. Oxford University Press.

Fletcher, Roger (1970). "A new approach to variable metric algorithms". Em: *The computer journal* 13.3, pp. 317–322.

Goldfarb, Donald (1970). "A family of variable-metric methods derived by variational means". Em: *Mathematics of computation* 24.109, pp. 23–26.

Greene, William H (1990). "A gamma-distributed stochastic frontier model". Em: *Journal of econometrics* 46.1-2, pp. 141–163.

— (2012a). *Econometric analysis*. 7ª ed. New York University: Prentice Hall.

— (2012b). *LIMDEP Reference Guide–Version 10, Plainview, New York: Econometric Software*.

Griffiths, William E e Hajargasht, Gholamreza (2016). "Some models for stochastic frontiers with endogeneity". Em: *Journal of Econometrics* 190.2, pp. 341–348.

Guan, Zhengfei et al. (2009). "Measuring excess capital capacity in agricultural production". Em: *American Journal of Agricultural Economics* 91.3, pp. 765–776.

Karakaplan, Mustafa e Kutlu, Levent (2015). "Handling endogeneity in stochastic frontier analysis". Em: *Available at SSRN 2607276*.

Karakaplan, Mustafa U (2017). "Fitting endogenous stochastic frontier models in Stata". Em: *The Stata Journal* 17.1, pp. 39–55.

Kumbhakar, Subal C e Lovell, CA Knox (2003). *Stochastic frontier analysis*. Cambridge university press.

Kutlu, Levent (2010). "Battese-Coelli estimator with endogenous regressors". Em: *Economics Letters* 109.2, pp. 79–81.

Meeusen, Wim e Den Broeck, Julien van (1977). "Efficiency estimation from Cobb-Douglas production functions with composed error". Em: *International economic review*, pp. 435–444.

Murphy, Kevin M e Topel, Robert H (1985). "Estimation and inference in two-step econometric models". Em: *Journal of Business & Economic Statistics* 3.4, pp. 370–379.

Mutter, Ryan L et al. (2013). "Investigating the impact of endogeneity on inefficiency estimates in the application of stochastic frontier analysis to nursing homes". Em: *Journal of Productivity Analysis* 39.2, pp. 101–110.

Shanno, David F (1970). "Conditioning of quasi-Newton methods for function minimization". Em: *Mathematics of computation* 24.111, pp. 647–656.

Souza, G da S, Gomes, Eliane Gonçalves e Alves, ER de A (2018). "Imperfeições de mercado e concentração de renda na produção agrícola." Em: *Área de Informação da Sede-Artigo em periódico indexado (ALICE)*.

Souza, Geraldo S, Gomes, Eliane G e Alves, Eliseu R A (2016). "Determinantes da dispersão da renda no meio rural brasileiro". Em: *Blucher Marine Engineering Proceedings* 2.1, pp. 173–184.

Stevenson, Rodney E (1980). "Likelihood functions for generalized stochastic frontier estimation". Em: *Journal of econometrics* 13.1, pp. 57–66.

Tran, Kien C e Tsionas, Efthymios G (2013). "GMM estimation of stochastic frontier model with endogenous regressors". Em: *Economics Letters* 118.1, pp. 233–236.

— (2015). "Endogeneity in stochastic frontier models: Copula approach without external instruments". Em: *Economics Letters* 133, pp. 85–88.

Wooldridge, Jeffrey M (2010). *Econometric analysis of cross section and panel data*. MIT press.

Apêndice A

Aplicação - Modelo normal/exponencial

Este apêndice ilustra os resultados da aplicação ao se considerar o modelo normal/exponencial, com suas estimativas obtidas em um ou dois estágios.

Como os resultados das regressões lineares por variáveis instrumentais da assistência técnica (*assistec*) e acesso a crédito (*finan*) estimadas em dois estágios são iguais para qualquer especificação de u_i, não foram repetidos aqui. Veja a Tabela 7.5.

As classificações normalizadas por região das medições de eficiência técnica preditas são apresentadas na Figura A.1. Observe como as eficiências diferem bastante de um modelo para o outro ao se especificar o modelo normal/exponencial.

Os resultados dos testes sobre a presença de endogeneidade são apresentados na Tabela A.4. Note que, em ambos os casos, rejeita-se a hipótese nula de exogeneidade.

Tabela A.1: Regressão linear por variáveis instrumentais da assistência técnica e acesso a crédito estimadas em um estágio assumindo que $u_i \sim Exp(\sigma_{ui})$.

Variável	Coeficiente	Erro padrão	z	P-valor	LI	LS
assistec						
constante	-0.331	0.037	-8.871	0.000	-0.404	-0.258
$\ln(terra)$	0.001	0.004	0.358	0.720	-0.006	0.009
$\ln(trab)$	0.006	0.004	1.814	0.070	-0.001	0.013
$\ln(cap)$	0.064	0.005	13.095	0.000	0.054	0.073
regiao$_{norte}$	0.216	0.016	13.452	0.000	0.184	0.247
regiao$_{nordeste}$	0.207	0.016	13.092	0.000	0.176	0.238
regiao$_{sudeste}$	0.168	0.014	12.010	0.000	0.141	0.196
regiao$_{sul}$	0.250	0.015	16.835	0.000	0.221	0.279
social	0.614	0.024	25.404	0.000	0.567	0.661
demo	0.081	0.029	2.786	0.005	0.024	0.138
ambi	0.070	0.035	2.013	0.044	0.002	0.139
gini	-0.545	0.034	-16.064	0.000	-0.612	-0.479
finan						
constante	-0.475	0.038	-12.553	0.000	-0.549	-0.401
$\ln(terra)$	0.015	0.004	3.764	0.000	0.007	0.023
$\ln(trab)$	-0.015	0.004	-4.143	0.000	-0.022	-0.008
$\ln(cap)$	0.134	0.005	27.435	0.000	0.125	0.144
regiao$_{norte}$	0.027	0.016	1.664	0.096	-0.005	0.058
regiao$_{nordeste}$	0.028	0.016	1.766	0.077	-0.003	0.059
regiao$_{sudeste}$	0.011	0.014	0.774	0.439	-0.017	0.038
regiao$_{sul}$	0.155	0.015	10.375	0.000	0.126	0.185
social	0.340	0.025	13.784	0.000	0.291	0.388
demo	-0.152	0.030	-5.040	0.000	-0.211	-0.093
ambi	-0.468	0.036	-12.934	0.000	-0.539	-0.397
gini	-0.277	0.035	-7.920	0.000	-0.346	-0.209

Tabela A.2: Parâmetros do modelo normal/exponencial estimados em um estágio.

Variável	Coeficiente	Erro padrão	z	P-valor	LI	LS
Fronteira						
constante	2.714	0.134	20.267	0.000	2.451	2.976
$\ln(terra)$	0.121	0.016	7.537	0.000	0.089	0.152
$\ln(trab)$	0.276	0.013	21.021	0.000	0.250	0.302
$\ln(cap)$	0.408	0.020	20.565	0.000	0.369	0.447
regiao$_{norte}$	0.255	0.060	4.215	0.000	0.136	0.373
regiao$_{nordeste}$	0.258	0.058	4.458	0.000	0.145	0.372
regiao$_{sudeste}$	0.420	0.051	8.262	0.000	0.320	0.519
regiao$_{sul}$	0.484	0.053	9.175	0.000	0.381	0.587
$\ln(\boldsymbol{\sigma_u^2})$						
constante	-1.054	0.570	-1.851	0.064	-2.171	0.062
assistec	2.915	0.390	7.483	0.000	2.151	3.679
finan	-4.902	0.372	-13.173	0.000	-5.632	-4.173
gini	-0.496	0.635	-0.780	0.435	-1.741	0.749
$\ln(\boldsymbol{\sigma_w^2})$						
constante	-1.060	0.025	-41.782	0.000	-1.110	-1.010
$\boldsymbol{\eta_{assistec}}$						
constante	1.004	0.089	11.298	0.000	0.830	1.178
$\boldsymbol{\eta_{finan}}$						
constante	-0.368	0.090	-4.091	0.000	-0.545	-0.192
σ_w^2	0.346	0.009	39.418	0.000	0.329	0.364

Tabela A.3: Parâmetros do modelo normal/exponencial estimados em dois estágios.

Variável	Coeficiente	Erro padrão	z	P-valor	LI	LS
Fronteira						
constante	1.985	0.252	7.865	0.000	1.490	2.480
$\ln(terra)$	0.162	0.018	9.049	0.000	0.127	0.196
$\ln(trab)$	0.215	0.017	12.903	0.000	0.182	0.248
$\ln(cap)$	0.492	0.024	20.875	0.000	0.446	0.538
regiao$_{norte}$	0.476	0.067	7.081	0.000	0.344	0.608
regiao$_{nordeste}$	0.343	0.063	5.434	0.000	0.219	0.466
regiao$_{sudeste}$	0.460	0.055	8.418	0.000	0.353	0.567
regiao$_{sul}$	0.534	0.056	9.457	0.000	0.423	0.645
$\boldsymbol{\ln(\sigma_u^2)}$						
constante	-6.027	2.982	-2.021	0.043	-11.872	-0.182
assistec	0.548	0.747	0.733	0.463	-0.916	2.012
finan	1.143	0.596	1.918	0.055	-0.025	2.312
gini	3.168	2.722	1.164	0.244	-2.166	8.503
$\boldsymbol{\ln(\sigma_w^2)}$						
constante	-0.960	0.048	-20.122	0.000	-1.053	-0.866
$\boldsymbol{\eta_{assistec}}$						
constante	0.915	0.094	9.718	0.000	0.731	1.100
$\boldsymbol{\eta_{finan}}$						
constante	1.034	0.085	12.207	0.000	0.868	1.200
σ_w^2	0.383	0.017	22.684	0.000	0.350	0.416

Tabela A.4: Teste de Wald e TRV para a presença de endogeneidade.

Abordagem	Wald	TRV	P-valor
MVIC	128	367	0.000
MVIL	340	229	0.000

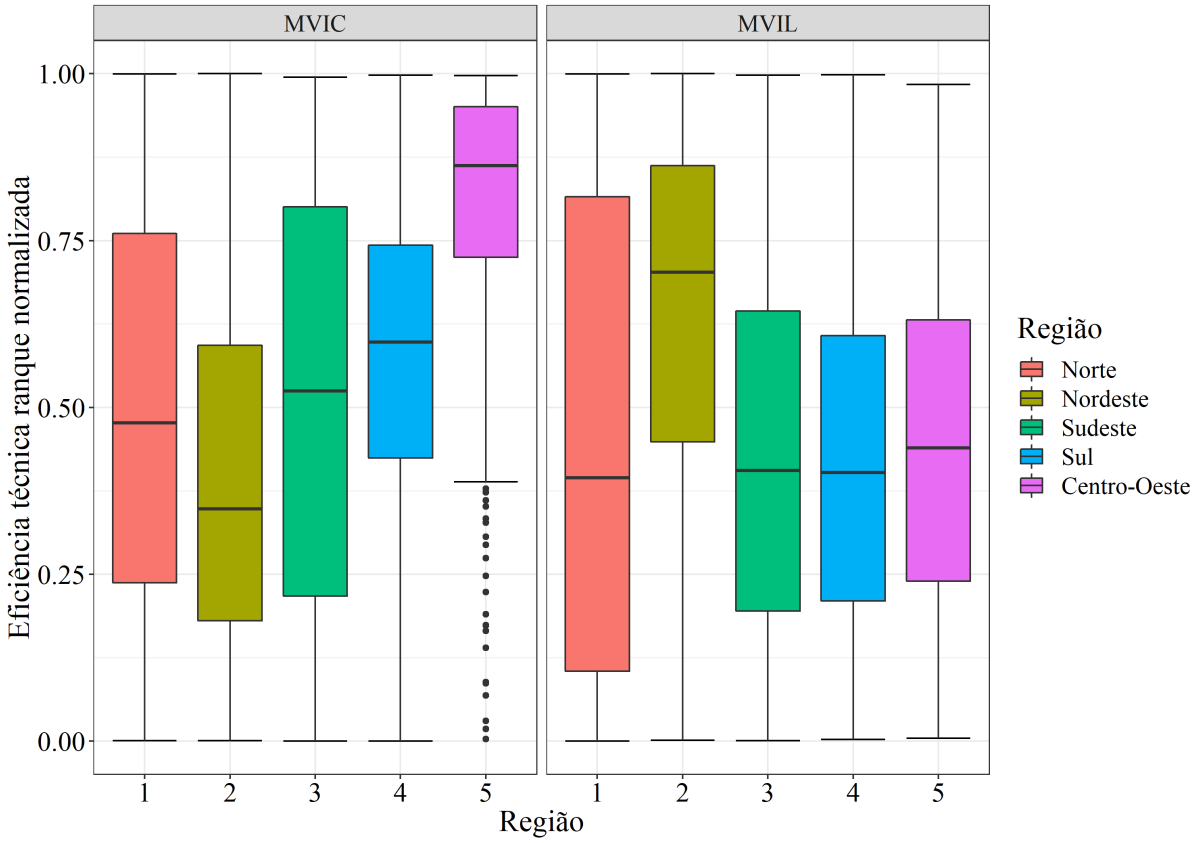

Figura A.1: *Box plot* normalizado das eficiências técnicas preditas pelos modelos normal/exponencial via abordagens MVIC e MVIL.

Apêndice B

Aplicação - Modelo normal/normal truncada

Este apêndice ilustra os resultados da aplicação ao se considerar o modelo normal/normal truncada, com suas estimativas obtidas em um ou dois estágios.

Como os resultados das regressões lineares por variáveis instrumentais da assistência técnica (*assistec*) e acesso a crédito (*finan*) estimadas em dois estágios são iguais para qualquer especificação de u_i, não foram repetidos aqui. Veja a Tabela 7.5.

As classificações normalizadas por região das medições de eficiência técnica preditas são apresentadas na Figura B.1. Como para o caso do modelo normal/seminormal, há semelhanças nas eficiências preditas pelos modelos normal/normal truncada via abordagens MVIC e MVIL.

Os resultados dos testes sobre a presença de endogeneidade são apresentados na Tabela B.4. Note que, em ambos os casos, rejeita-se a hipótese nula de exogeneidade.

Tabela B.1: Regressão linear por variáveis instrumentais da assistência técnica e acesso a crédito estimadas em um estágio assumindo que $u_i \sim \mathcal{N}^+(\mu_i, \sigma_u^2)$.

Variável	Coeficiente	Erro padrão	z	P-valor	LI	LS
assistec						
constante	-0.283	0.036	-7.903	0.000	-0.354	-0.213
$\ln(terra)$	0.002	0.004	0.571	0.568	-0.005	0.010
$\ln(trab)$	0.003	0.003	0.998	0.318	-0.003	0.010
$\ln(cap)$	0.079	0.005	16.761	0.000	0.070	0.088
regiao$_{norte}$	0.055	0.015	3.585	0.000	0.025	0.085
regiao$_{nordeste}$	0.046	0.015	3.028	0.002	0.016	0.075
regiao$_{sudeste}$	0.064	0.013	4.769	0.000	0.038	0.090
regiao$_{sul}$	0.163	0.014	11.413	0.000	0.135	0.191
social	0.470	0.023	20.819	0.000	0.426	0.514
demo	0.048	0.027	1.798	0.072	-0.004	0.100
ambi	-0.050	0.034	-1.445	0.149	-0.117	0.018
gini	-0.426	0.029	-14.661	0.000	-0.483	-0.369
finan						
constante	-0.514	0.036	-14.197	0.000	-0.585	-0.443
$\ln(terra)$	0.026	0.004	6.442	0.000	0.018	0.034
$\ln(trab)$	-0.008	0.003	-2.316	0.021	-0.014	-0.001
$\ln(cap)$	0.128	0.005	26.701	0.000	0.119	0.137
regiao$_{norte}$	-0.073	0.015	-4.715	0.000	-0.103	-0.043
regiao$_{nordeste}$	-0.082	0.015	-5.355	0.000	-0.112	-0.052
regiao$_{sudeste}$	-0.059	0.014	-4.310	0.000	-0.085	-0.032
regiao$_{sul}$	0.107	0.015	7.352	0.000	0.078	0.135
social	0.175	0.023	7.637	0.000	0.130	0.220
demo	-0.175	0.030	-5.844	0.000	-0.234	-0.117
ambi	-0.447	0.034	-13.275	0.000	-0.513	-0.381
gini	-0.134	0.030	-4.535	0.000	-0.192	-0.076

Tabela B.2: Parâmetros do modelo normal/normal truncada estimados em um estágio.

Variável	Coeficiente	Erro padrão	z	P-valor	LI	LS
Fronteira						
constante	10.498	0.112	94.149	0.000	10.279	10.717
$\ln(terra)$	0.055	0.016	3.380	0.001	0.023	0.087
$\ln(trab)$	0.114	0.013	8.576	0.000	0.088	0.141
$\ln(cap)$	0.113	0.040	2.807	0.005	0.034	0.192
regiao$_{norte}$	0.145	0.060	2.416	0.016	0.027	0.262
regiao$_{nordeste}$	0.105	0.058	1.820	0.069	-0.008	0.218
regiao$_{sudeste}$	-0.031	0.056	-0.564	0.572	-0.141	0.078
regiao$_{sul}$	-0.357	0.071	-5.040	0.000	-0.495	-0.218
$\boldsymbol{\mu}$						
constante	8.143	0.120	68.019	0.000	7.909	8.378
assistec	-2.215	0.168	-13.160	0.000	-2.545	-1.885
finan	-1.528	0.279	-5.476	0.000	-2.075	-0.981
gini	-4.288	0.140	-30.558	0.000	-4.563	-4.013
$\ln(\boldsymbol{\sigma_u^2})$						
constante	-4.866	0.002	-2706.075	0.000	-4.869	-4.862
$\ln(\boldsymbol{\sigma_w^2})$						
constante	-1.220	0.021	-59.211	0.000	-1.260	-1.179
$\boldsymbol{\eta_{assistec}}$						
constante	-1.537	0.175	-8.772	0.000	-1.880	-1.193
$\boldsymbol{\eta_{finan}}$						
constante	-1.087	0.284	-3.831	0.000	-1.642	-0.531
σ_u^2	0.008	0.000	556.173	0.000	0.008	0.008
σ_w^2	0.295	0.006	48.551	0.000	0.283	0.307
σ^2	0.303	0.021	14.303	0.000	0.262	0.345
λ	0.162	0.018	8.960	0.000	0.126	0.197

Tabela B.3: Parâmetros do modelo normal/normal truncada estimados em dois estágios.

Variável	Coeficiente	Erro padrão	z	P-valor	LI	LS
Fronteira						
constante	10.308	0.119	86.474	0.000	10.074	10.541
$\ln(terra)$	0.058	0.019	3.072	0.002	0.021	0.095
$\ln(trab)$	0.114	0.016	7.176	0.000	0.083	0.145
$\ln(cap)$	0.131	0.044	3.002	0.003	0.046	0.217
regiao$_{norte}$	0.139	0.065	2.156	0.031	0.013	0.266
regiao$_{nordeste}$	0.100	0.061	1.654	0.098	-0.019	0.219
regiao$_{sudeste}$	-0.027	0.057	-0.474	0.635	-0.139	0.085
regiao$_{sul}$	-0.331	0.069	-4.782	0.000	-0.467	-0.195
$\boldsymbol{\mu}$						
constante	8.043	0.121	66.461	0.000	7.806	8.280
assistec	-2.172	0.169	-12.869	0.000	-2.503	-1.841
finan	-1.426	0.274	-5.194	0.000	-1.964	-0.888
gini	-4.249	0.157	-27.007	0.000	-4.557	-3.940
$\ln(\boldsymbol{\sigma_u^2})$						
constante	-4.582	0.001	-4602.257	0.000	-4.584	-4.580
$\ln(\boldsymbol{\sigma_w^2})$						
constante	-1.226	0.025	-48.782	0.000	-1.275	-1.176
$\boldsymbol{\eta_{assistec}}$						
constante	-1.490	0.175	-8.524	0.000	-1.833	-1.147
$\boldsymbol{\eta_{finan}}$						
constante	-0.981	0.278	-3.535	0.000	-1.525	-0.437
σ_u^2	0.010	0.000	1205.013	0.000	0.010	0.010
σ_w^2	0.294	0.006	48.205	0.000	0.282	0.306
σ^2	0.304	0.021	14.150	0.000	0.262	0.346
λ	0.187	0.004	50.181	0.000	0.179	0.194

Tabela B.4: Teste de Wald e TRV para a presença de endogeneidade.

Abordagem	Wald	TRV	P-valor
MVIC	120	1959	0.000
MVIL	114	1746	0.000

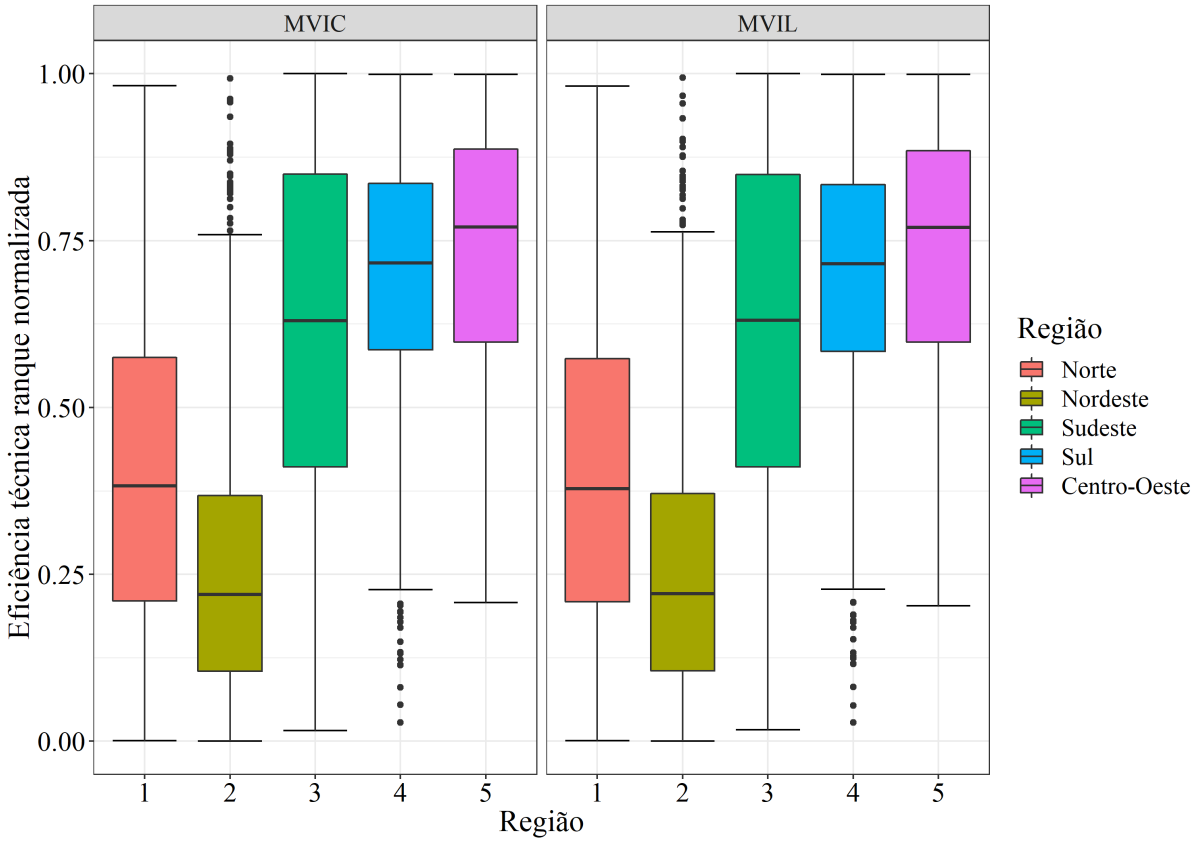

Figura B.1: *Box plot* normalizado das eficiências técnicas preditas pelos modelos normal/normal truncada via abordagens MVIC e MVIL.

Apêndice C

Entrada de dados no R

```
> rm(list = ls())
> setwd("C:/Users/Kessys/Desktop/Dissertação/Scripts")
> load("dados.RData")
> reg <- factor(dados$regiao, levels = c(5, 1, 2, 3, 4))
> source("SFprod.R")
>
> mod.h1 <- SF.prod(model = "half1s",
+                   fr.form  = ly ~ lxterra + lxtrab + lxcap + reg,
+                   s2u.form = ~ assistec + finan + gini,
+                   s2w.form = ~ 1,
+                   end.form = cbind(assistec, finan) ~ lxterra + lxtrab + lxcap + reg +
    social + demo + ambi + gini,
+                   data = dados)
initial  value 5419.217567
iter   2 value 4427.035383
iter   3 value 3401.014888
...
iter  60 value 874.773765
final  value 874.773765
converged

> mod.h1
$table
           Coefficient     Std.Err           z      P.value   Lower.limit  Upper.limit
constant   -0.260465726 0.035792092  -7.2771863 3.408385e-13  -0.330616938 -0.190314514
lxterra     0.004931063 0.003952956   1.2474367 2.122374e-01  -0.002816589  0.012678715
lxtrab      0.010298701 0.003354037   3.0705391 2.136727e-03   0.003724910  0.016872493
lxcap       0.082043281 0.004743959  17.2942637 0.000000e+00   0.072745292  0.091341270
```

```
reg1         0.047914524 0.015310726   3.1294744 1.751193e-03   0.017906053  0.077922995
reg2         0.048342926 0.015100193   3.2014774 1.367248e-03   0.018747092  0.077938760
reg3         0.071062465 0.013429433   5.2915464 1.212864e-07   0.044741261  0.097383669
reg4         0.165414800 0.014331416  11.5421110 0.000000e+00   0.137325740  0.193503861
social       0.405833456 0.022858521  17.7541429 0.000000e+00   0.361031577  0.450635334
demo        -0.015830602 0.028264190  -0.5600939 5.754154e-01  -0.071227397  0.039566193
ambi         0.040233833 0.034091589   1.1801689 2.379330e-01  -0.026584453  0.107052120
gini        -0.579275416 0.030675740 -18.8838288 0.000000e+00  -0.639398761 -0.519152071
constant.1  -0.508170485 0.036456294 -13.9391703 0.000000e+00  -0.579623508 -0.436717463
lxterra.1    0.027315415 0.003998493   6.8314278 8.407275e-12   0.019478513  0.035152317
lxtrab.1    -0.004705936 0.003407409  -1.3810890 1.672516e-01  -0.011384335  0.001972464
lxcap.1      0.128958117 0.004798721  26.8734348 0.000000e+00   0.119552797  0.138363437
reg1.1      -0.075636649 0.015485090  -4.8844823 1.037010e-06  -0.105986869 -0.045286430
reg2.1      -0.080141084 0.015291933  -5.2407424 1.599318e-07  -0.110112723 -0.050169446
reg3.1      -0.057489371 0.013612623  -4.2232397 2.408155e-05  -0.084169622 -0.030809119
reg4.1       0.104278521 0.014547726   7.1680291 7.609469e-13   0.075765503  0.132791539
social.1     0.155713099 0.023629402   6.5898029 4.404099e-11   0.109400323  0.202025875
demo.1      -0.221999128 0.029303687  -7.5758088 3.574918e-14  -0.279433298 -0.164564957
ambi.1      -0.397513744 0.035578418 -11.1728899 0.000000e+00  -0.467246162 -0.327781327
gini.1      -0.196524548 0.032284451  -6.0872818 1.148438e-09  -0.259800908 -0.133248187
constant.2   2.477120650 0.124777464  19.8523080 0.000000e+00   2.232561314  2.721679987
lxterra.2    0.110463770 0.014460570   7.6389638 2.198242e-14   0.082121574  0.138805967
lxtrab.2     0.206309481 0.011496967  17.9446879 0.000000e+00   0.183775840  0.228843122
lxcap.2      0.509498093 0.017583090  28.9765952 0.000000e+00   0.475035869  0.543960317
reg1.2       0.066021692 0.054956270   1.2013496 2.296156e-01  -0.041690618  0.173734002
reg2.2       0.129015693 0.053193512   2.4254028 1.529141e-02   0.024758325  0.233273062
reg3.2       0.241634809 0.046372419   5.2107441 1.880848e-07   0.150746538  0.332523081
reg4.2       0.335258751 0.047876610   7.0025582 2.513323e-12   0.241422319  0.429095183
constant.3   5.637215449 0.201044521  28.0396373 0.000000e+00   5.243175428  6.031255470
assistec     0.930905808 0.525480912   1.7715312 7.647242e-02  -0.099017853  1.960829469
finan       -2.931351182 0.579237593  -5.0607060 4.177068e-07  -4.066636004 -1.796066361
gini.2      -9.937942741 0.298732915 -33.2669828 0.000000e+00 -10.523448494 -9.352436987
constant.4  -1.077703565 0.022543224 -47.8060978 0.000000e+00  -1.121887472 -1.033519658
eta.1        0.889708382 0.079483930  11.1935631 0.000000e+00   0.733922742  1.045494022
eta.2        0.200051085 0.080528016   2.4842421 1.298275e-02   0.042219075  0.357883096
s2w          0.340376281 0.007673179  44.3592277 0.000000e+00   0.325337127  0.355415435

$summary.ef
V1
Min.   :0.3448
1st Qu.:0.7943
Median :0.8601
Mean   :0.8404
```

```
3rd Qu.:0.9059
Max.   :0.9763

$sd.ef
[1] 0.08936858

$value
[1] -874.7738

$AIC
[1] 1827.548

$BIC
[1] 2081.444

$`cor pearson IV`
        assistec.IV  finan.IV
assistec   0.8159787 0.7548839
finan      0.7455923 0.8070101

$`cor spearman IV`
        assistec.IV  finan.IV
assistec   0.8159234 0.7624692
finan      0.7468320 0.8189413

$cor.pearson
[1] 0.8791935

$cor.spearman
[1] 0.8856089

$`LR chisq test`
statistic      P-value
[1,]  341.9781 5.500406e-75

$`Wald chisq test`
statistic      P-value
[1,]  200.6754 2.653902e-44

$bias
[1] 0.1910636

$RMSE
```

```
[1] 1.135411

$`sample size`
[1] 4965

$`estimated parameters`
[1] 39

> head(mod.h1$efficiency,2)
[,1]
1 0.8348427
2 0.8230292

> head(mod.h1$error,2)
[,1]
1 -0.1358302
2 -0.5815880

> head(mod.h1$fitted.y_without.corretion,2)
[,1]
1 10.35010
2 10.47475

> head(mod.h1$fitted.y_with.corretion,2)
[,1]
1 10.20869
2 10.52470



> mod.h2 <- SF.prod(model = "half2s",
+                   fr.form  = ly ~ lxterra + lxtrab + lxcap + reg,
+                   s2u.form = ~ assistec + finan + gini,
+                   s2w.form = ~ 1,
+                   end.form = cbind(assistec, finan) ~ lxterra + lxtrab + lxcap + reg +
    social + demo + ambi + gini,
+                   data = dados)
initial  value 9071.047380
iter   2 value 8078.865195
iter   3 value 7052.743487
...
iter  32 value 4543.961971
final  value 4543.961971
converged
```

```
>
$reg_IV
Response assistec :

Call:
lm(formula = assistec ~ lxterra + lxtrab + lxcap + reg + social +
demo + ambi + gini, data = data)

Residuals:
Min       1Q   Median       3Q      Max
-0.60807 -0.11472 -0.00733  0.10742  0.73351

Coefficients:
            Estimate   Std. Error t value Pr(>|t|)
constant    -0.292583   0.035909  -8.148 4.65e-16 ***
lxterra      0.003401   0.003937   0.864 0.387741
lxtrab       0.003867   0.003314   1.167 0.243390
lxcap        0.078877   0.004719  16.716  < 2e-16 ***
reg1         0.054932   0.015267   3.598 0.000324 ***
reg2         0.047122   0.015086   3.124 0.001797 **
reg3         0.059204   0.013398   4.419 1.01e-05 ***
reg4         0.155140   0.014339  10.819  < 2e-16 ***
social       0.487451   0.022421  21.740  < 2e-16 ***
demo        -0.003171   0.029041  -0.109 0.913062
ambi        -0.017765   0.034868  -0.510 0.610423
gini        -0.426269   0.029111 -14.643  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1681 on 4953 degrees of freedom
Multiple R-squared:  0.6686,  Adjusted R-squared:  0.6679
F-statistic: 908.4 on 11 and 4953 DF,  p-value: < 2.2e-16

Response finan :

Call:
lm(formula = finan ~ lxterra + lxtrab + lxcap + reg + social +
demo + ambi + gini, data = data)

Residuals:
Min       1Q   Median       3Q      Max
-1.23510 -0.07927  0.01717  0.10585  0.68120
```

```
Coefficients:
            Estimate   Std. Error t value Pr(>|t|)
constant    -0.521363   0.036427 -14.313  < 2e-16 ***
lxterra      0.026687   0.003994   6.682 2.62e-11 ***
lxtrab      -0.007348   0.003362  -2.186   0.0289 *
lxcap        0.127658   0.004787  26.669  < 2e-16 ***
reg1        -0.072754   0.015487  -4.698 2.70e-06 ***
reg2        -0.080642   0.015304  -5.269 1.43e-07 ***
reg3        -0.062359   0.013591  -4.588 4.58e-06 ***
reg4         0.100060   0.014546   6.879 6.78e-12 ***
social       0.189234   0.022745   8.320  < 2e-16 ***
demo        -0.216800   0.029459  -7.359 2.15e-13 ***
ambi        -0.421335   0.035370 -11.912  < 2e-16 ***
gini        -0.133683   0.029531  -4.527 6.12e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1706 on 4953 degrees of freedom
Multiple R-squared:  0.6517,  Adjusted R-squared:  0.651
F-statistic: 842.7 on 11 and 4953 DF,  p-value: < 2.2e-16

$table
            Coefficient     Std.Err           z      P.value   Lower.limit  Upper.limit
constant     2.55227812 0.125691533  20.3058873 0.000000e+00   2.30592724   2.7986290
lxterra      0.11081235 0.014733277   7.5212290 5.417888e-14   0.08193566   0.1396890
lxtrab       0.20198172 0.012013659  16.8126725 0.000000e+00   0.17843538   0.2255281
lxcap        0.50719634 0.018028512  28.1330123 0.000000e+00   0.47186110   0.5425316
reg1         0.06178397 0.055007445   1.1231928 2.613556e-01  -0.04602864   0.1695966
reg2         0.13073348 0.053242913   2.4554157 1.407218e-02   0.02637929   0.2350877
reg3         0.23087906 0.046222076   4.9949956 5.883711e-07   0.14028546   0.3214727
reg4         0.31920732 0.047387806   6.7360646 1.627343e-11   0.22632893   0.4120857
constant.1   6.33669654 0.594318067  10.6621301 0.000000e+00   5.17185454   7.5015385
assistec    -0.18183052 0.577537436  -0.3148376 7.528849e-01  -1.31378309   0.9501221
finan       -2.21491060 0.616611954  -3.5920656 3.280672e-04  -3.42344783  -1.0063734
gini       -10.44125027 0.840651690 -12.4204238 0.000000e+00 -12.08889731  -8.7936032
constant.2  -1.06518964 0.023447293 -45.4291087 0.000000e+00  -1.11114549  -1.0192338
residuo.1    0.66420096 0.079685575   8.3352722 0.000000e+00   0.50802010   0.8203818
residuo.2    0.23826935 0.079807914   2.9855354 2.830824e-03   0.08184871   0.3946900
s2w          0.34466249 0.007844294  43.9379842 0.000000e+00   0.32928795   0.3600370

$summary.ef
V1
Min.   :0.2723
```

```
1st Qu.:0.7855
Median :0.8530
Mean   :0.8336
3rd Qu.:0.9025
Max.   :0.9691

$sd.ef
[1] 0.0923872

$value
[1] -892.1322

$AIC
[1] 1814.264

$BIC
[1] 1911.917

$`cor pearson IV`
        assistec.IV  finan.IV
assistec   0.8176718 0.7556542
finan      0.7460742 0.8073055

$`cor spearman IV`
        assistec.IV  finan.IV
assistec   0.8161624 0.7626196
finan      0.7466482 0.8186686

$cor.pearson
[1] 0.8792244

$cor.spearman
[1] 0.8854699

$`LR chisq test`
statistic      P-value
[1,]  207.6339 8.182035e-46

$`Wald chisq test`
statistic      P-value
[1,]  149.3292 3.746095e-33

$bias
```

```
[1] 0.201474

$RMSE
[1] 1.12561

$`sample size`
[1] 4965

$`estimated parameters`
[1] 15

> head(mod.h2$efficiency,2)
[,1]
1 0.8053387
2 0.8199075

> head(mod.h2$error,2)
[,1]
1 -0.1937421
2 -0.5859230

> head(mod.h2$fitted.y_without.corretion,2)
[,1]
1 10.36468
2 10.48685

> head(mod.h2$fitted.y_with.corretion,2)
[,1]
1 10.26660
2 10.52904
```