Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Laís Raiane Miguel Amaral

Estimação em dois estágios de um modelo de fronteira estocástica de produção aplicado a dados do Sistema Nacional de Avaliação da Educação Básica

Brasília
2019

Laís Raiane Miguel Amaral

Estimação em dois estágios de um modelo de fronteira estocástica de produção aplicado a dados do Sistema Nacional de Avaliação da Educação Básica

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade de Brasília como requisito para a obtenção do título de Mestre em Estatística.

Área de Concentração: Probabilidade e Estatística
Orientador: Prof. Bernardo Borba de Andrade
Coorientador: Prof. Geraldo da Silva e Souza

Brasília
2019

Aos meus amados Gustavo e Cecília.

Agradecimentos

A Deus, que me impulsiona a viver com realismo, dedicação e coragem.

Ao Gustavo, meu amado esposo, por toda cumplicidade, companheirismo e amor. Por viver absolutamente tudo ao meu lado.

À minha amada filha Cecília, meu grande presente desse período do mestrado. Você trouxe a força e a determinação necessárias para enfrentar o final deste curso. Sua alegria é o que me motiva diariamente.

À minha família, por todo apoio, amor e união.

À minha grande amiga Camila, amiga que a estatística me trouxe, grande companheira na vida e também nos estudos. Às minhas amigas Marília e Luana, pela torcida e amor de sempre.

Aos inúmeros colegas com quem compartilhei aulas e trabalhos, em especial Bruno e Kessys, pela generosidade, amizade e companheirismo.

Aos professores do departamento de Estatística da UnB, em especial à professora Cira, pela dedicação e humanidade com as quais exerce a docência.

Aos professores Geraldo e Bernardo, pela orientação desta dissertação.

Por fim, aos colegas do Inep, por todo apoio ao longo desse período.

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

"*A coisa mais extraordinária do mundo é um homem comum, uma mulher comum e seus filhos comuns*"

G. K. Chesterton

Resumo

Esta dissertação apresenta uma revisão bibliográfica dos modelos mais comuns de fronteira estocástica de produção na literatura (modelo normal-seminormal, modelo normal-normal truncada e modelo normal-exponencial). Discutimos o uso da correção de Murphy-Topel (2002) para matriz de variância-covariância em estimação de dois estágios, em que uma probabilidade estimada (primeiro estágio) é usada como variável explicativa na estimação de um modelo de SFA no segundo estágio. Ilustramos a metodologia com uma base de dados educacionais no software R. A unidade observada é a escola e a variável resposta é a média da nota obtida pela escola na Prova Brasil em Português e Matemática. A base de dados foi construída a partir dos microdados do Censo Escolar e do Saeb, que são instrumentos fornecidos pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anı´sio Teixeira (Inep).

Palavras-chave: Análise de Fronteira Estocástica de Produção (SFA), estimação de máxima verossimilhança em dois estágios, correção de Murphy-Topel, eficiência técnica, R, Censo Escolar, Sistema Nacional de Avaliação da Educação Básica (Saeb).

Abstract

This monograph provides a bibliographical review of the most common stochastic frontier models in the literature (normal-half normal model, normal-truncated normal model and normal-exponential model). We discuss the use of Murphy-Topel's (2002) correction for variance-covariance matrix in two-stage estimation where an estimated probability (first-stage) is used as explanatory variable in the estimation of an SFA model in the second stage. We illustrate the methodology with educational data in the R system. The data consist of schools as observational units and the response variable is the average grade obtained by the school in Prova Brasil for Portuguese and Mathematics. The database was constructed from the Educational Census microdata and from Saeb provided by the Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep).

Key words: Stochastic Frontier Analysis (SFA), two-stage maximum likelihood estimation, Murphy-Topel correction, technical efficiency, R, Censo Escolar, Sistema Nacional de Avaliação da Educação Básica (Saeb).

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

1.1 Considerações iniciais

A qualidade da educação brasileira é tema recorrente em estudos e discussões no país. No âmbito da Companhia de Planejamento do Distrito Federal (Codeplan), Rosa et al. (2016) apresentaram a situação do Distrito Federal no contexto nacional em relação à eficiência técnica das escolas públicas brasileiras. Para este estudo, Rosa et al. (2016) utilizaram dados da Avaliação Nacional do Rendimento Escolar (Anresc), avaliação que faz parte do Sistema de Avaliação da Educação Básica (Saeb), popularmente conhecida como Prova Brasil, e do Censo Escolar da Educação Básica. Ambos conjuntos de dados são de responsabilidade do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep).

Rosa et al. (2016) produziram um modelo de fronteira estocástica de produção, técnica que foi introduzida por Meeusen e van Den Broeck (1977) e Aigner et al. (1977), para tratar problemas da Teoria de Firma, mas que vem sendo amplamente utilizada em outras áreas, como saúde e educação. A análise de fronteira estocástica assume que cada firma produz potencialmente menos do que poderia devido a um grau de ineficiência. Os modelos de fronteira estocástica de produção são tipicamente estimados por máxima verossimilhança ou por mínimos quadrados corrigidos e a consistência de ambos estimadores depende da exogeneidade das variáveis que compõe os fatores de produção ou que modelam a ineficiência técnica.

Esta dissertação propõe estender o estudo de Rosa et al. (2016) ao utilizar o mesmo conjunto de dados para estimar uma fronteira estocástica de produção em dois estágios, incluindo na análise indicadores disponibilizados pelo Inep a partir de 2014. Os indicadores

da educação básica podem ser usados para reduzir o número de variáveis utilizadas por Rosa et al. (2016) na construção da fronteira estocástica.

A estimação em dois estágios é um procedimento simples. A partir de um modelo auxiliar, os componentes não observados são substituídos por suas estimativas ou valores preditos. Criticamente, em muitas aplicações esses valores são tratados como se fossem conhecidos, em razão da estimação e inferência no modelo do segundo estágio, que usualmente é o modelo de interesse.

Procedimentos em dois estágios produzem estimativas consistentes dos parâmetros do modelo do segundo estágio sob condições gerais. Porém, a estimativa dos erros padrão e estatísticas de testes relacionados a eles são incorretas. Murphy e Topel (2002) propuseram uma correção para a matriz de variância-covariância do modelo estimado no segundo estágio, que também será objeto de estudo nesta dissertação.

1.2 Dados

Nesta dissertação foram utilizados os dados do Censo Escolar da Educação Básica e do Saeb. A média das notas de português e matemática obtidas na Prova Brasil pelos alunos de determinada escola é o produto do modelo. Recursos físicos são considerados como insumos. E para ineficiência, variáveis que captam o contexto socioeconômico em que os alunos da escola estão inseridos, e aspectos da gestão escolar e dos docentes.

Foram abordadas a estimação em dois estágios e a correção de Murphy-Topel para matriz de variância-covariância do modelo do segundo estágio.

O modelo estimado no segundo estágio indica que uma quantidade maior de recursos físicos está associada a melhores resultados. Em relação à ineficiência, o histórico escolar do aluno tem uma importante relevância. Em relação aos professores, a regularidade do corpo docente na escola se destacou. Já a variável de complexidade de gestão da escolar apresentou um resultado inesperado.

1.3 Esboço do trabalho

Este estudo foi desenvolvido da seguinte forma: o capítulo 2 apresenta a revisão bibliográfica dos modelos de fronteira estocástica de produção. O capítulo 3 explica o método da máxima verossimilhança em dois estágios e o estimador de Murphy-Topel para variância

do modelo do segundo estágio. No capítulo 4 é relatada a base de dados. No capítulo 5 é realizada a aplicação da estimação em dois estágios e a correção de Murphy-Topel nos dados educacionais. Por fim, no capítulo 6 é apresentada a conclusão e sugestão de trabalhos futuros.

Capítulo 2

Fronteira estocástica de produção

Um relevante problema a ser estudado na teoria econômica consiste em como medir a performance de diferentes firmas em relação à maneira como convertem insumos em produtos considerando produtividade e eficiência. A eficiência econômica se refere a dois componentes: eficiência técnica, que mede a habilidade da firma em obter o máximo de produto a partir dos insumos; e alocação eficiente, que mede a habilidade da firma em usar a proporção ótima de insumos em relação aos seus custos.

Diferentes fatores podem fazer com que uma firma não obtenha a possibilidade máxima de produção admitida pela tecnologia atual, por isso surgiram as medidas de ineficiência técnica. Existem diversos métodos, paramétricos e não paramétricos, para medir a performance de firmas. Em relação às abordagens paramétricas, três principais processos são sugeridos: 1) considerar que a distância entre a produção de uma firma e a produção ótima se deve a erros aleatórios bilaterais, que representam as variações ocasionadas por fatores que não podem ser controlados pela firma, o que corresponde a um modelo clássico de regressão; 2) considerar que a distância se deve à ineficiência da firma, variável aleatória estritamente positiva, o que corresponde a uma fronteira determinística; 3) considerar que a distância é resultante de dois termos: tanto dos erros aleatórios bilaterais, quanto da ineficiência (como na fronteira determinística, variável aleatória estritamente positiva), o que corresponde a uma fronteira estocástica de produção (SFA, do inglês *Stochastic Frontier Analysis*).

As abordagens paramétricas são apresentadas na tabela (2.1), na qual $v \in \mathbb{R}$ é o erro aleatório bilateral e $u \in \mathbb{R}_+$ é a ineficiência. É possível observar que tanto nas especificações aditivas, quanto nas multiplicativas, o erro aleatório v pode elevar ou diminuir a produção. Por outro lado, a ineficiência u sempre irá provocar uma queda na produção.

Tabela 2.1 - Abordagens paramétricas

Abordagem	Aditivo	Multiplicativo
Regressão	$y = f(x;\beta) + v$	$y = f(x;\beta)\exp(v)$
Determinística	$y = f(x;\beta) - u$	$y = f(x;\beta)\exp(-u)$
Estocástica	$y = f(x;\beta) + v - u$	$y = f(x;\beta)\exp(v)\exp(-u)$

2.1 Modelo clássico de regressão

Conforme visto anteriormente, a fronteira de produção para um conjunto de firmas pode ser estimada como um modelo de regressão via mínimos quadrados ordinários (MQO). Nesta abordagem, assume-se que os erros em relação à fronteira são aleatórios, normalmente distribuídos e simétricos em torno de zero.

$$y_i = f(x_i;\beta) + v_i, \qquad v_i \sim N(0,\sigma^2).$$

Suponha uma situação hipotética em que deseja-se estimar a fronteira de produção de 190 escolas de determinado território, em que o produto das escolas será dado pela nota média obtida pelos seus alunos em um exame de proficiência e o insumo será o número de tarefas de casa que eles fizeram no ano em que o teste de proficiência foi aplicado.

Considere que o conjunto de i escolas, $i = 1, ..., 190$, tem função de produção do tipo Cobb-Douglas, isto é, cada firma i terá produção igual a

$$y_i = \beta_0 + \beta_1 x_i,$$

em que y_i representa o logaritmo da produção de cada escola i, x_i representa o logaritmo do insumo e β_0 e β_1 os parâmetros desconhecidos.

A figura (2.1) ilustra esse modelo para o caso hipotético. O modelo estimado foi

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 2,02 + 0,65 x_i.$$

Uma vez que a soma dos resíduos é igual a zero, a fronteira de produção estimada se localiza no centro de todas as observações. Como se deseja estimar o máximo produto possível a partir do insumo dado, o modelo clássico de regressão se mostra inadequado, pois aproximadamente metade das observações se encontram acima da fronteira estimada, o que é claramente uma contradição.

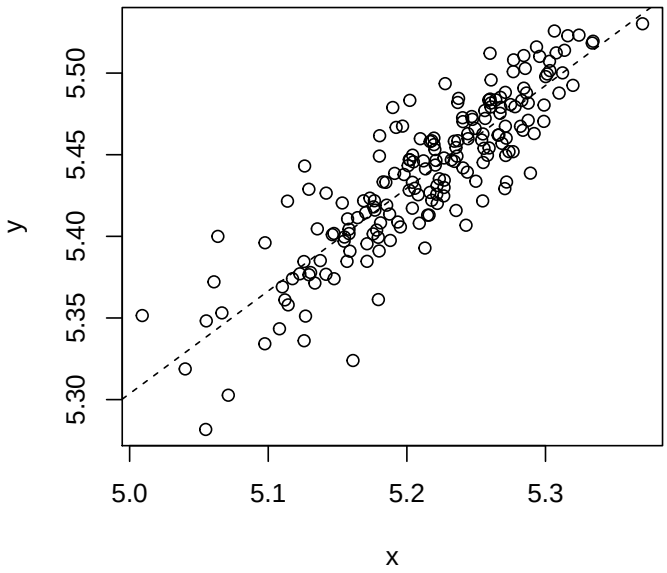

Figura 2.1: Modelo clássico de regressão

2.2 Fronteira determinística

Por outro lado, pode-se assumir que todos os desvios são resultados da ineficiência das firmas, supondo o modelo (2.1).

$$y_i = f(x_i; \beta) - u_i, \qquad u_i \sim H, \tag{2.1}$$

em que H é alguma distribuição de probabilidade com suporte apenas em $\mathbb{R}_+$. Na literatura, as distribuições mais comumente utilizadas para H são: exponencial, gama, seminormal e normal truncada.

Bogetoft e Otto (2010) afirmam que uma das formas mais amplamente utilizadas para estimar a fronteira determinística é via mínimos quadrados ordinários corrigidos (MQOC), o que se obtém em dois passos.

O primeiro passo é calcular os estimadores dos parâmetros via MQO. A equação (2.2) mostra que o estimador do intercepto será tendencioso.

Considere o modelo de fronteira determinística com uma variável regressora:

$$y_i = \beta_0 + \beta_1 x_i - u_i,$$

o estimador de MQO para β_1 será dado por

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

e sua esperança será dada por

$$\begin{aligned}
\mathrm{E}\left(\hat{\beta}_1\right) &= \mathrm{E}\left(\frac{\sum_{i=1}^{n} y_i (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right) \\
&= \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sum_{i=1}^{n} (x_i - \bar{x}) \mathrm{E}(y_i) \\
&= \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sum_{i=1}^{n} (x_i - \bar{x}) (\beta_0 + \beta_1 - \mathrm{E}(u)) \\
&= \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \left[\beta_0 \sum_{i=1}^{n} (x_i - \bar{x}) + \beta_1 \sum_{i=1}^{n} x_i (x_i - \bar{x}) - \mathrm{E}(u) \sum_{i=1}^{n} (x_i - \bar{x})\right] \\
&= \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \beta_1 \sum_{i=1}^{n} x_i (x_i - \bar{x}) \\
&= \frac{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)} \beta_1 \\
&= \beta_1.
\end{aligned}$$

O estimador de β_1 é não tendencioso. Por outro lado, o estimador de mínimos quadrados para β_0 será dado por

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

e sua esperança será

$$\begin{aligned}
\mathrm{E}(\beta_0) &= \mathrm{E}\left(\bar{y} - \hat{\beta}_1 \bar{x}\right) \\
&= \frac{1}{n} \mathrm{E}\left(\sum_{i=1}^{n} y_i\right) - \bar{x} \mathrm{E}\left(\hat{\beta}_1\right) \\
&= \frac{1}{n} \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i - \mathrm{E}(u)) - \bar{x}\beta_1 \\
&= \beta_0 + \bar{x}\beta_1 - \mathrm{E}(u) - \bar{x}\beta_1 \\
&= \beta_0 - \mathrm{E}(u). \qquad (2.2)
\end{aligned}$$

Como $\mathrm{E}(u) \geq 0$, o estimador de MQO do intercepto do modelo de fronteira determinística será tendencioso quando $\mathrm{E}(u) > 0$.

O segundo passo é encontrar a menor correção possível para $\hat{\beta}_0$ que assegure que todas as observações fiquem abaixo da fronteira de produção, isto é, ajustar β com o erro máximo.

$$\tilde{\beta_0} = \beta_0 + \max_i \{\hat{u}_i\}, \tag{2.3}$$

em que $\hat{u}_i$ são os resíduos gerados pelos estimadores de MQO. Eles podem ser

$$-\tilde{u}_i = \hat{u}_i - \max_i \{\hat{u}_i\}. \tag{2.4}$$

Os resíduos de MQOC, $\tilde{u}_i$, são não-negativos, com pelo menos um igual a zero. A técnica de MQOC é de fácil implementação, porém a fronteira de produção determinística estimada é paralela à regressão clássica, uma vez que apenas o intercepto de MQO é corrigido. Isto implica que a estrutura da tecnologia de produção determinística é a mesma estrutura da tecnologia de produção de tendência central. Esta é uma indesejável propriedade restritiva do procedimento de MQOC, pois a estrutura da tecnologia de produção deveria ser diferente da tecnologia de produção central, em que os produtores são menos eficientes do que os melhores produtores (Kumbhakar e Lovell (2003)). Isto é, a fronteira de MQOC não necessariamente vincula os dados na melhor fronteira possível.

A figura (2.2) ilustra a fronteira determinística (linha pontilhada) e o modelo clássico (linha tracejada) para o caso hipotético. O estimador do intercepto determinístico foi dado pela equação (2.5) e o modelo de regressão determinístico pela equação (2.6):

$$\tilde{\beta_0} = \hat{\beta}_0 + \max_i \{\hat{u}_i\} \tag{2.5}$$

$$\tilde{y_i} = 2,08 + 0,65x_i. \tag{2.6}$$

Como sugerido, todas as observações ficam abaixo da fronteira determinística estimada.

2.3 Fronteira estocástica de produção

A análise de fronteira estocástica assume que cada firma potencialmente produz menos do que poderia, apresentando dois erros: um erro aleatório bilateral e um grau de ineficiência. Especificamente,

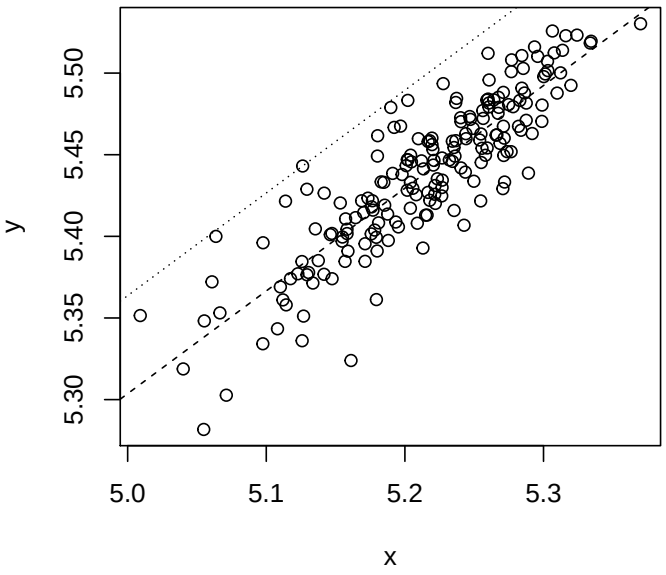

Figura 2.2: Fronteira determinística

$$\log(y) = \log(f(x_i;\beta)) + v_i - u_i, \qquad v_i \sim N(0,\sigma^2), \qquad u_i \sim H, \tag{2.7}$$

v_i é o erro aleatório simétrico, u_i é uma variável aleatória associada com a ineficiência técnica e H é alguma distribuição de probabilidade com suporte apenas em $\mathbb{R}_+$. Na literatura, as distribuições mais comumente utilizadas para H são: exponencial, gama, seminormal e normal truncada. Os modelos de fronteira estocástica com essas distribuições serão detalhados na seção 2.4.

Considerando que o lado direito da equação (2.7) é composto por duas variáveis aleatórias, os métodos de estimação dos parâmetros da fronteira estocástica de produção são sustentados por suposições sobre estas duas variáveis aleatórias.

É comum assumir que cada v_i é distribuído independentemente de cada u_i, e que ambos os erros são não correlacionados com as variáveis explicativas. Além disso, supõe-se que

$$\begin{aligned}
&\mathrm{E}(v_i) = 0,\\
&\mathrm{E}(v_i^2) = \sigma_v^2,\\
&\mathrm{E}(v_i v_j) = 0, \text{ para todo } i \neq j,\\
&\mathrm{E}(u_i^2) = \sigma_u^2,\\
&\mathrm{E}(u_i u_j) = 0, \text{ para todo } i \neq j.
\end{aligned}$$

Sob essas suposições, é possível obter os estimadores dos parâmetros via mínimos quadrados ou mínimos quadrados corrigidos, porém os os problemas descritos anteriormente permaneceriam.

Uma solução mais adequada é construir suposições em relação às distribuições de probabilidade dos componentes do erro e estimar os parâmetros da fronteira estocástica de produção usando o método da máxima verossimilhança. Este método é preferível devido às propriedades assintóticas dos estimadores de máxima verossimilhança para grandes amostras.

Para a situação hipotética apresentada nas seções anteriores, o modelo de fronteira estocástica foi dado pela equação (2.8):

$$\breve{y} = 2,05 + 0,65x_i \tag{2.8}$$

A figura (2.3) representa a fronteira estocástica de produção estimada pelo método da máxima verossimilhança em que $u_i \sim N_+(0, \sigma_u^2)$ (linha sólida), a fronteira determinística (linha pontilhada) e o modelo clássico (linha tracejada). Nota-se que ainda existem observações acima da fronteira estocástica de produção, porém estas são atribuídas ao erro aleatório.

2.3.1 Estimação da eficiência técnica de cada firma

Um dos pontos de interesse da análise de fronteira estocástica de produção é a estimação da eficiência técnica de cada firma. A forma mais comum de mensurar tal eficiência é calcular a razão entre a produção observada e sua correspondente função de produção estocástica sem a presença de ineficiência:

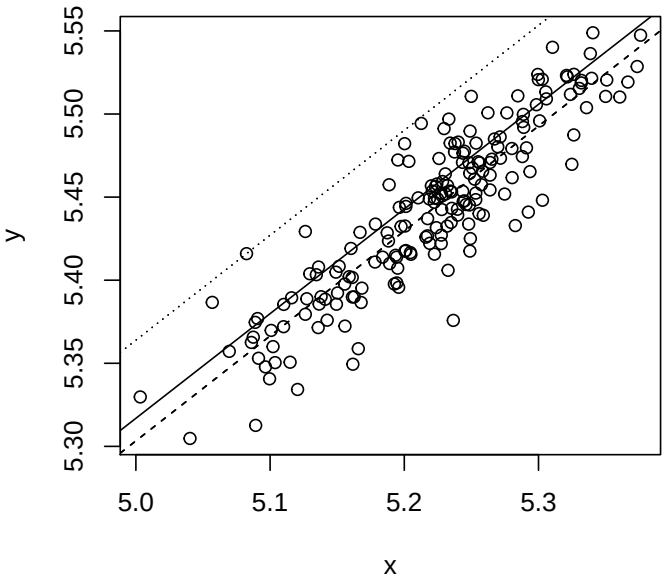

Figura 2.3: Fronteira estocástica de produção

$$\theta_i = \frac{y_i}{f(x_i;\beta)\exp(v_i)} = \frac{f(x_i;\beta)\exp(v_i)\exp(-u_i)}{f(x_i;\beta)\exp(v_i)} = \exp(-u_i). \tag{2.9}$$

Na equação (2.9), θ_i assume valores entre zero e um. Se $\theta_i = 1$ a firma apresenta uma produção eficiente. Para valores menores do que um, essa medida indica o quão distante a produção de uma firma está em relação à produção que poderia obter, utilizando os mesmos insumos, sem ineficiência.

2.3.2 Principais tecnologias na função de produção

Na literatura de análise de eficiência, diversos tipos de função podem ser utilizados para modelar a fronteira de produção (Coelli et al. (2005)). Alguns exemplos estão dispostos na tabela (2.2).

2.4 Modelos de Fronteira Estocástica

Inicialmente, foram propostas as distribuições seminormal (Aigner et al. (1977)) e exponencial (Meeusen e van Den Broeck (1977)) como densidade de u para os modelos de

Tabela 2.2 - Principais tecnologias de produção

Linear	$y = \beta_0 + \sum_{n=1}^{N} \beta_n x_n$
Cobb-Douglas	$y = \beta_0 \prod_{n=1}^{N} x_n^{\beta_n}$
Quadrática	$y = \beta_0 + \sum_{n=1}^{N} \beta_n x_n + \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \beta_{nm} x_n x_m$
Quadrática normalizada	$y = \beta_0 + \sum_{n=1}^{N-1} \beta_n \frac{x_n}{x_N} + \frac{1}{2} \sum_{n=1}^{N-1} \sum_{m=1}^{N-1} \beta_{nm} \frac{x_n}{x_N} \frac{x_m}{x_N}$
Translog	$y = \exp\left(\beta_0 + \sum_{n=1}^{N} \beta_n \log(x_n) + \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \beta_{nm} \log(x_n) \log(x_m)\right)$
Leontief generalizada	$y = \sum_{n=1}^{N} \sum_{m=1}^{N} \beta_{nm} \left(x_n x_m\right)^{\frac{1}{2}}$
Elasticidade de substituição constante (CES)	$y = \beta_0 \left(\sum_{n=1}^{N} \beta_n x_n^{\gamma}\right)^{\frac{1}{\gamma}}$

fronteira estocástica de produção. Uma vez que essas distribuições tem moda igual a zero, adotá-las significa implicitamente assumir que a maior parte das firmas tem um nível de eficiência próximo de um, pois $e^0 = 1$.

Diante dessa problemática, posteriormente, essas especificações originais foram generalizadas pela distribuição normal truncada (Stevenson (1980)) e pela distribuição gama (Beckers e Hammond (1987); Greene (1990)). Pois são distribuições mais flexíveis, com suporte no eixo positivo, admitem uma moda diferente de zero para u e, consequentemente, permitem que a eficiência da firma não seja necessariamente um.

No caso da normal truncada, o custo dessa flexibilização é um parâmetro extra, o ponto de truncamento μ, que pode inflacionar o erro padrão dos outros parâmetros, uma vez que a curvatura da função de verossimilhança é achatada e, com isso, o determinante da matriz de Fischer aumenta. Esse modelo também pode apresentar problemas de convergência no processo de estimação iterativo.

No caso do modelo obtido com a distribuição gama, a dificuldade consiste no fato de que a função de verossimilhança correspondente a ele não tem uma forma fechada. Isso significa que para utilizá-la é necessário aproximar o valor da função de verossimilhança a partir de métodos numéricos, o que costuma ser difícil ou, em alguns casos, até impossível.

2.4.1 O modelo normal-seminormal

Considere o modelo de fronteira estocástica de produção no qual o termo de ineficiência assume uma distribuição seminormal.

$$y = f(x_i;\beta) + v_i - u_i, \qquad v_i \sim N(0,\sigma_v^2), \qquad u_i \sim N_+(0,\sigma_u^2).$$

A fim de determinar, pelo método da máxima verossimilhança, os valores dos parâmetros β desconhecidos e de σ_v^2 e σ_u^2, é necessário conhecer a densidade conjunta do termo de erro, cuja distribuição é uma convolução de uma distribuição normal, v, com uma distribuição seminormal, u,

$$\varepsilon = v - u. \tag{2.10}$$

Quando o modelo é estimado para encontrar β, σ_v^2 e σ_u^2 é possível calcular facilmente os termos de erro:

$$\varepsilon_i = v_i - u_i = y_i - f(x_i;\beta),$$

porém não é possível obter diretamente os componentes v_i e u_i.

Apesar de existir interesse em estimar o termo de erro, no contexto de fronteira estocástica de produção, o interesse maior recai sobre a estimação das ineficiências individuais, isto é, a estimação de u_i, $i = 1, ..., n$. Esta é uma importante questão, que será abordada a partir da distribuição de ϵ.

2.4.1.1 A função de verossimilhança

Para utilizar o método da máxima verossimilhança, inicialmente, é necessário obter a função de verossimilhança.

A função densidade para uma única observação do termo de erro, v, é a distribuição normal:

$$f_v(v) = \frac{1}{\sqrt{2\pi\sigma_v^2}} e^{\frac{-1}{2}\frac{v^2}{\sigma_v^2}}. \tag{2.11}$$

E a densidade para o termo de ineficiência u é a distribuição seminormal, ou seja, a distribuição normal truncada no zero,

$$f_u(u)\begin{cases}\frac{2}{\sqrt{2\pi\sigma_u^2}}e^{\frac{-1}{2}\frac{u^2}{\sigma_u^2}}, & \text{para todo u } \geq 0\\ 0, & \text{para todo u } < 0.\end{cases}$$

O erro total dado pela equação (2.10) é a soma de v e $-u$, com isso, a distribuição de ε é a convolução da distribuição de v e de $-u$, que será obtida por

$$f_\varepsilon(\varepsilon) = \int_{-\infty}^{\infty} f_u(u)f_v(\varepsilon+u)du = \int_0^{\infty} f_u(u)f_v(\varepsilon+u)du.$$

Para facilitar o cálculo da integral, inicialmente, manipula-se o produto das densidades

$$\begin{aligned} f_v(\varepsilon+u)f_u(u) &= \frac{1}{\sqrt{2\pi\sigma_v^2}}e^{-\frac{1}{2}\frac{(\varepsilon+u)^2}{\sigma_v^2}}\frac{2}{\sqrt{2\pi\sigma_u^2}}e^{-\frac{1}{2}\frac{u^2}{\sigma_u^2}} \\ &= \frac{1}{\pi\sqrt{\sigma_u^2\sigma_v^2}}e^{-\frac{1}{2}\frac{u^2}{\sigma_u^2}-\frac{1}{2}\frac{(\varepsilon+u)^2}{\sigma_v^2}} \\ &= \frac{1}{\pi\sqrt{\sigma_u^2\sigma_v^2}}e^{-\frac{1}{2}\frac{(\sigma_u^2+\sigma_v^2)u^2+2\sigma_u^2\varepsilon u+\sigma_u^2\varepsilon^2}{\sigma_u^2\sigma_v^2}}. \end{aligned}$$

Agora é possível calcular a integral, que envolve os seguintes passos: [1]

$$\begin{aligned} f_\varepsilon(\varepsilon) &= \int_0^{\infty} f_v(\varepsilon+u)f_u(u)du \\ &= \frac{1}{\pi\sqrt{\sigma_u^2\sigma_v^2}}\int_0^{\infty} e^{-\frac{1}{2}\frac{(\sigma_u^2+\sigma_v^2)u^2+2\sigma_u^2\varepsilon u+\sigma_u^2\varepsilon^2}{\sigma_u^2\sigma_v^2}}du \\ &= \frac{1}{\pi\sqrt{\sigma_u^2\sigma_v^2}}\sqrt{\frac{\pi}{2}}\frac{1}{\sqrt{\frac{1}{\sigma_u^2}+\frac{1}{\sigma_v^2}}}\left(1-\text{erf}\left(\frac{\varepsilon}{\sqrt{2}\sqrt{\frac{1}{\sigma_u^2}+\frac{1}{\sigma_v^2}}\sigma_v^2}\right)e^{-\frac{1}{2}\frac{\varepsilon^2}{(\sigma_u^2+\sigma_v^2)}}\right) \\ &= \frac{1}{\sqrt{2\pi}\sqrt{\sigma_u^2+\sigma_v^2}}\left(1-\text{erf}\left(\frac{\varepsilon}{\sqrt{2}\sqrt{\sigma_u^2+\sigma_v^2}}\sqrt{\frac{\sigma_u^2}{\sigma_v^2}}\right)\right)e^{-\frac{1}{2}\frac{\varepsilon^2}{(\sigma_u^2+\sigma_v^2)}}. \end{aligned}$$

Faça $\sigma^2 = \sigma_v^2 + \sigma_u^2$ e $\lambda = \sqrt{\frac{\sigma_u^2}{\sigma_v^2}}$.

[1] A função erro $\text{erf}(x) = \frac{2}{\pi}\int_0^x e^{-t^2}dt$, utilizada nos próximos cálculos, tem a seguinte propriedade: $\text{erf}(-x) = -\text{erf}(x)$. O relacionamento desta função com a distribuição normal é dado por $\Phi(x) - \frac{1}{2} = \frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{1}{2}t^2}dt = \frac{1}{2}\text{erf}\left(\frac{x}{\sqrt{2}}\right)$, isto é, $\Phi(x) = \frac{1}{2}\left(1+\text{erf}\left(\frac{x}{\sqrt{2}}\right)\right)$.

$$\begin{aligned} f_\varepsilon(\varepsilon) &= \frac{1}{\sqrt{2\pi\sigma^2}}\left(1-\text{erf}\left(\frac{\varepsilon}{\sqrt{2}\sqrt{\sigma^2}}\lambda\right)\right)e^{-\frac{1}{2}\frac{\varepsilon^2}{\sigma^2}} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\left(1+\text{erf}\left(-\frac{\lambda\varepsilon}{\sqrt{2\sigma^2}}\right)\right)e^{-\frac{1}{2}\frac{\varepsilon^2}{\sigma^2}} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}2\Phi\left(-\frac{\lambda\varepsilon}{\sqrt{\sigma^2}}\right)e^{-\frac{1}{2}\frac{\varepsilon^2}{\sigma^2}} \\ &= \frac{\sqrt{2}}{\sqrt{\pi\sigma^2}}\Phi\left(-\frac{\lambda\varepsilon}{\sqrt{\sigma^2}}\right)e^{-\frac{1}{2}\frac{\varepsilon^2}{\sigma^2}}, \end{aligned}$$

em que Φ é a função distribuição da normal padrão com média zero e variância um. Quando o parâmetro λ é zero, não há impacto originado por diferenças de ineficiências, por outro lado, quando λ é expressivo, isso se deve a diferenças na ineficiência e não no erro aleatório.

O log dessa densidade é

$$\log f_\varepsilon(\varepsilon) = -\frac{1}{2}\log\left(\frac{\pi}{2}\right) - \frac{1}{2}\log\sigma^2 + \log\Phi\left(-\frac{\varepsilon\lambda}{\sqrt{\sigma^2}}\right) - \frac{1}{2}\frac{\varepsilon^2}{\sigma^2}.$$

Considerando o caso de n observações independentes, a densidade conjunta é

$$f\left(\varepsilon_1,...,\varepsilon_n\right) = \prod_{i=1}^{n} f_\varepsilon(\varepsilon_i)$$

e o log da densidade conjunta é

$$\begin{aligned} \log f\left(\varepsilon_1,...,\varepsilon_n\right) &= \sum_{i=1}^{n}\log f_\epsilon(\varepsilon_i) \\ &= -\frac{1}{2}n\log\left(\frac{\pi}{2}\right) - \frac{1}{2}n\log\sigma^2 + \sum_{i=1}^{n}\log\Phi\left(-\frac{\lambda\varepsilon_i}{\sqrt{\sigma^2}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\varepsilon_i^2 \\ &= \text{constante} - n\log(\sigma) + \sum_{i=1}^{n}\log\Phi\left(-\frac{\lambda\varepsilon_i}{\sigma}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\varepsilon_i^2. \end{aligned} \tag{2.12}$$

É possível reescrever a equação (2.12) para enfatizar que o termo de erro ϵ depende do parâmetro β. Com isso, a função de log-verossimilhança é dada por

$$\begin{aligned} l\left(\beta,\sigma^2,\lambda\right) &= \log f_\varepsilon\left(\varepsilon_1(\beta),...,\varepsilon_n(\beta);\sigma^2;\lambda\right) \\ &= \log f_\varepsilon\left(y_1 - f\left(x_1;\beta\right),...,y_n - f\left(x_n;\beta\right);\sigma^2;\lambda\right) \\ &= \text{constante} - n\log\sigma + \sum_{i=1}^{n}\log\Phi\left(-\frac{\lambda\left(y_i - f\left(x_i;\beta\right)\right)}{\sqrt{\sigma^2}}\right) \\ &\quad -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - f\left(x_i;\beta\right)\right)^2. \end{aligned} \tag{2.13}$$

2.4.1.2 Método numérico

A solução para o problema de maximização da equação (2.13) pode ser resolvido com a solução das equações de verossimilhança,

$$\frac{\partial l(\beta)}{\partial \beta_j} = 0, \quad j = 1,...,m.$$

Infelizmente, como a função de verossimilhança para o modelo de fronteira estocástica não é linear em seus parâmetros e não existe fórmula fechada para obtenção dos parâmetros, a estimação é realizada por um processo iterativo de otimização. Isto pode ser feito a partir da expansão de 1ª ordem de Taylor na equação da log-verossimilhança para obter

$$0 = \frac{\partial l(\hat{\beta})}{\partial \beta} \simeq \frac{\partial l(\beta^0)}{\partial \beta} + \left(\hat{\beta} - \beta^0\right)\frac{\partial^2 l\left(\beta^0\right)}{\partial \beta^2},$$

tal que

$$\hat{\beta} \simeq \beta^0 - \left(\frac{\partial^2 l(\beta^0)}{\partial \beta^2}\right)^{-1}\frac{\partial l(\beta^0)}{\partial \beta}.$$

A equação (2.14) pode ser utilizada em um processo iterativo para solucionar β:

$$\beta^{n+1} = \beta^n - \left(\frac{\partial^2 l(\beta^n)}{\partial \beta^2}\right)^{-1}\frac{\partial l(\beta^n)}{\partial \beta}. \tag{2.14}$$

Inicialmente, escolha uma valor para β^0 e repita a fórmula (2.14) substituindo β^n pelo novo valor β^{n+1} calculado. Repita o processo até que o valor de β^{n+1} não difira de β^n, isto é, $|\beta^{n+1} - \beta^n| < \varepsilon$ para algum $\varepsilon > 0$. Usualmente, assume-se $\varepsilon = 10^{-4}$. Se o valor de β for muito pequeno, pode ser utilizado o critério $\frac{\left|\beta^{n+1}-\beta^n\right|}{\beta^n+1} < \varepsilon$. Este é o método de Newton. Como chute inicial para β^0, podem ser utilizados os parâmetros estimados por mínimos quadrados ou por mínimos quadrados corrigidos.

2.4.1.3 Variância da eficiência

O processo descrito gera estimativas para λ e para σ^2, porém, é necessário estimar também σ_u^2 e σ_v^2. Isto pode ser feito de maneira simples resolvendo as equações $\sigma^2 = \sigma_u^2 + \sigma_v^2$ e $\lambda = \sqrt{\frac{\sigma_u^2}{\sigma_v^2}}$:

$$\lambda = \sqrt{\frac{\sigma_u^2}{\sigma_v^2}} \Rightarrow \lambda^2 = \frac{\sigma_u^2}{\sigma_v^2} \Rightarrow \sigma_u^2 = \lambda^2 \sigma_v^2.$$

Agora,

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

Por fim,

$$\sigma_u^2 = \lambda^2 \sigma_v^2 = \frac{\lambda^2}{1 + \lambda^2}\sigma^2.$$

2.4.1.4 Eficiência específica

A eficiência específica de uma firma depende de u tanto no modelo aditivo, quanto no multiplicativo. No modelo multiplicativo, a eficiência depende apenas de u, enquanto no modelo aditivo, a eficiência depende também do máximo produto esperado. A eficiência específica será dada por (2.15) para o modelo aditivo e por (2.16) para o modelo multiplicativo.

$$\theta_i = (x_i, y_i) = \frac{f\left(x_i, \hat{\beta}\right) - \hat{u}_i}{f\left(x_i, \hat{\beta}\right)} = 1 - \frac{\hat{u}_i}{f\left(x_i, \hat{\beta}\right)} \tag{2.15}$$

$$\theta_i = \exp\left(-\hat{u}_i\right) \tag{2.16}$$

Independente do modelo utilizado, é preciso estimar u_i para calcular a eficiência específica. Infelizmente, encontrar $\hat{u}_i$ não é simples. Por outro lado, após estimar os parâmetros, é possível calcular também o erro total

$$\hat{\varepsilon_i} = y_i - f\left(x_i; \hat{\beta}\right), \quad i = 1, ..., n. \tag{2.17}$$

Como o erro total é dado por $\varepsilon_i = v_i - u_i$, o seu estimador carrega alguma informação de u_i. Se $\varepsilon_i > 0$, há chances de u_i não ser muito grande, pois $\mathrm{E}(v_i) = 0$ e $u_i \geq 0$, o que

sugere que a firma i é relativamente eficiente. Por outro lado, se $\varepsilon_i < 0$, então u_i tende a ser grande, sugerindo que a firma i é relativamente ineficiente.

Com isso, uma solução é estimar u_i em função de ε_i, a partir da probabilidade condicional de u_i dado ε_i. Omita por um momento os índices, a densidade conjunta de v e u é o produto das densidades individuais, uma vez que as variáveis são independentes, logo $f_{u,v}(u,v) = f_v(v)f_u(u)$. Substituindo v por $\varepsilon + u$, obtém-se $f_{u,\varepsilon}(u,\varepsilon) = f_v(\varepsilon+u)f_u(u)$. Então, usando o Teorema de Bayes, a densidade condicional de u dado ε é

$$\begin{aligned} f(u \mid \varepsilon) &= \frac{f_v(\varepsilon+u)\, f_u(u)}{f_\varepsilon(\varepsilon)} \\ &= \frac{1}{\sqrt{2\pi}\sigma_*}\exp\left\{-\frac{(u-\mu_*)^2}{2\sigma_*^2}\right\} \Big/ \left[1-\Phi\left(-\frac{\mu_*}{\sigma_*}\right)\right], \end{aligned}$$

em que,

$$\begin{aligned} \mu_* &= -\varepsilon\frac{\sigma_u^2}{\sigma^2} = -\varepsilon\frac{\lambda^2}{1+\lambda^2} = -\varepsilon\gamma, \\ \sigma_* &= \sqrt{\frac{\sigma_u^2\sigma_v^2}{\sigma^2}} = \frac{\lambda}{1+\lambda^2}\sigma = \sqrt{\gamma(1-\gamma)\sigma^2}. \end{aligned}$$

$f(u \mid \varepsilon)$ segue uma distribuição $N^+(\mu_*, \sigma_*^2)$, ao substituir os valores estimados para ε, σ^2 e λ na equação (2.18), a média dessa distribuição pode ser utilizada como estimador para u_i.

$$E(u \mid \varepsilon) = \mu_* + \sigma_*\frac{\phi(\mu_*/\sigma_*)}{\Phi(\mu_*/\sigma_*)}, \tag{2.18}$$

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

Uma vez que $\mathrm{E}(\theta) = \mathrm{E}(e^{-u})$ geralmente não é igual a $e^{-\mathrm{E}(u)}$, o estimador (2.19) foi também proposto. Este estimador é ótimo em relação à minimização do erro quadrático médio (Bogetoft e Otto (2010)).

$$\theta = \mathrm{E}(\exp\{-u \mid \varepsilon\}) = \left[\frac{\Phi(\mu_*/\sigma_* - \sigma_*)}{\Phi(\mu_*/\sigma_*)}\right]\exp\left\{-\mu_* + \frac{1}{2}\sigma_*^2\right\}. \tag{2.19}$$

2.4.2 O modelo normal-exponencial

Considere agora o modelo de fronteira estocástica de produção no qual o termo de ineficiência assume uma distribuição exponencial, isto é:

$$f(u) = \frac{1}{\sigma_u} e^{-u\frac{1}{\sigma_u}}. \tag{2.20}$$

A densidade conjunta de v e u será dada pelo produto das densidades (2.11) e (2.20), pois as variáveis são independentes:

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

Logo, a densidade conjunta de u e ε é

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

E a densidade marginal de ϵ é

$$f(\varepsilon) = \frac{1}{\sigma_u}\Phi\left(-\frac{\varepsilon}{\sigma_v} - \frac{\sigma_v}{\sigma_u}\right) \exp\left\{\frac{\varepsilon}{\sigma_u} + \frac{\sigma_v^2}{2\sigma_u^2}\right\}.$$

Em Kumbhakar e Lovell (2003) é apresentada a função de log-verossimilhança do modelo de fronteira estocástica cuja ineficiência é distribuída exponencialmente:

$$\log f(\varepsilon_1, ..., \varepsilon_n) = \text{constante} - n\log\sigma_u + \sum_{i=1}^{n} \log\Phi(-A) + \sum_{i=1}^{n} \frac{\varepsilon_i}{\sigma_u}, \tag{2.21}$$

em que $A = -\tilde{\mu}/\sigma_v$ e $\tilde{\mu} = -\varepsilon - (\sigma_v^2/\sigma_u)$. A equação (2.21) pode ser utilizada para se obter os estimadores de máxima verossimilhança de todos os parâmetros.

Como no modelo seminormal, os estimadores pontuais das eficiências técnicas podem ser obtidos a partir da média da distribuição condicional de u dado ε. A distribuição condicional $f(u \mid \varepsilon)$ é dada por

$$f(u \mid \varepsilon) = \frac{1}{\sqrt{2\pi}\sigma_v\Phi(-\tilde{\mu}/\sigma_v)} \exp\left\{-\frac{(u-\tilde{\mu})^2}{2\sigma^2}\right\}.$$

$f(u \mid \varepsilon)$ segue uma distribuição $N_+(\tilde{\mu}, \sigma_v^2)$, com média

$$\mathrm{E}(u \mid \varepsilon) = \tilde{\mu} + \sigma_v\left[\frac{\phi(-\tilde{\mu}/\sigma_v)}{\Phi(\tilde{\mu}/\sigma_v)}\right] = \sigma_v\left[\frac{\phi(A)}{\Phi(-A)} - A\right],$$

em que $\phi(\cdot)$ e $\Phi(\cdot)$ novamente representam a densidade e a distribuição acumulada da normal padrão.

2.4.3 O modelo normal-normal truncado

Uma outra possibilidade de modelo de fronteira estocástica de produção é aquele em que a ineficiência segue uma distribuição normal truncada, isto é,

$$f(u) = \frac{1}{\sqrt{2\pi}\sigma_u \Phi\left(-\mu/\sigma_u\right)} \exp\left\{-\frac{(u-\mu)^2}{2\sigma_u^2}\right\}, \tag{2.22}$$

na qual μ é a moda da distribuição normal truncada, $\Phi(\cdot)$ é a função de distribuição acumulada da normal padrão. Logo, $f(u)$ é a densidade de uma variável normalmente distribuída com possibilidade de média μ diferente de zero. Se $\mu = 0$ a densidade da equação (2.22) coincide com a densidade seminormal apresentada na equação (2.12).

A densidade conjunta de u e v é o produto das suas densidades individuais:

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

E a densidade conjunta de u e ε é

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

Por fim, a densidade marginal de ε é

$$\begin{aligned} f(\varepsilon) &= \frac{1}{\sqrt{2\pi}\sigma \Phi\left(-\mu/\sigma_u\right)} \Phi\left(\frac{\mu}{\sigma\lambda} - \frac{\varepsilon\lambda}{\sigma}\right) \exp\left\{-\frac{(\varepsilon+u)^2}{2\sigma^2}\right\} \\ &= \frac{1}{\sigma}\phi\left(\frac{\varepsilon+\mu}{\sigma}\right) \Phi\left(\frac{\mu}{\sigma\lambda} - \frac{\varepsilon\lambda}{\sigma}\right) \left[\Phi\left(-\frac{\mu}{\sigma_u}\right)\right]^{-1}, \end{aligned} \tag{2.23}$$

na qual $\sigma = \left(\sigma_u^2 + \sigma_v^2\right)^{1/2}$ e $\lambda = \sigma_u/\sigma_v$.

A equação (2.23) é assimetricamente distribuída, com média e variância dadas por

$$\mathrm{E}(\varepsilon) = -\mathrm{E}(u) = -\frac{\mu a}{2} - \frac{\sigma_u a}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\mu}{\sigma_u}\right)^2\right\} \text{ e}$$

$$\mathrm{V}(\varepsilon) = \mu^2 \frac{a}{2}\left(1 - \frac{a}{2}\right) + \frac{a}{2}\left(\frac{\pi - a}{\pi}\right)\sigma_u^2 + \sigma_v^2,$$

em que $a = \left[\Phi\left(-\mu/\sigma_u\right)\right]^{-1}$. A equação do erro (2.23) tem três parâmetros, um de locação, μ, e dois de dispersão, σ_u e σ_v.

Kumbhakar e Lovell (2003) apresentaram a função de log-verossimilhança para um conjunto de n firmas

$$\begin{aligned} \log f\left(\varepsilon_1, ..., \varepsilon_n\right) &= \text{constante} - n\log\sigma - n\log\Phi\left(-\frac{\mu}{\sigma_u}\right) \\ &\quad + \sum_{i=1}^{n} \log\Phi\left(\frac{\mu}{\sigma\lambda} - \frac{\lambda}{\sigma}\right) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{\varepsilon+\mu}{\sigma}\right)^2, \end{aligned}$$

em que $\sigma_u = \lambda\sigma/\sqrt{1+\lambda^2}$.

A distribuição conjunta $f(u \mid \varepsilon)$ é dada por

$$f(u \mid \varepsilon) = \frac{1}{\sqrt{2\pi}\sigma_* \left[1 - \Phi\left(-\tilde{\mu}/\sigma_*\right)\right]} \exp\left\{-\frac{(u-\tilde{\mu})^2}{2\sigma_*^2}\right\}.$$

$f(u \mid \varepsilon)$ segue uma distribuição $N_+(\tilde{\mu}, \sigma_*^2)$, em que $\tilde{\mu} = \left(-\sigma_u^2\varepsilon + \mu\sigma_v^2\right)/\sigma^2$ e $\sigma_*^2 = \sigma_u^2\sigma_v^2/\sigma^2$. Assim como nas seções anteriores, a média da distribuição condicional pode ser utilizada para estimar a eficiência técnica individual de cada firma, ao substituir os parâmetros calculados em

$$\mathrm{E}(u \mid \varepsilon) = \sigma_* \left[\frac{\tilde{\mu}}{\sigma_*} + \frac{\phi\left(\tilde{\mu}/\sigma_*\right)}{1 - \Phi\left(-\tilde{\mu/\sigma_*}\right)}\right]$$

A eficiência técnica específica pode ser calculada também por

$$\theta = \mathrm{E}\left(\exp\left\{-u \mid \varepsilon\right\}\right) = \frac{1 - \Phi\left[\sigma_* - \left(\tilde{\mu}/\sigma_*\right)\right]}{1 - \Phi\left(-\tilde{\mu/\sigma_*}\right)} \exp\left\{-\tilde{\mu} + \frac{1}{2}\sigma_*^2\right\}$$

2.4.4 O modelo normal-gama

Assim como o modelo seminormal pode ser generalizado assumindo que u segue uma distribuição normal truncada, o modelo exponencial pode ser generalizado assumindo que u segue uma distribuição gama. A distribuição gama generaliza o parâmetro da exponencial adicionando um parâmetro a ser estimado, e assim fornece uma representação mais flexível do padrão de eficiência técnica nos dados.

Considere que $u \sim Gama(\alpha, \lambda)$, com $\mathrm{E}(u) = \frac{\alpha}{\lambda}$, a densidade de u será dada por

$$f(u) = \frac{\lambda^\alpha}{\Gamma(\alpha)} u^{\alpha-1} e^{-\lambda u}, \quad \alpha, \lambda > 0.$$

Para $\alpha = 1$ a densidade gama se torna a densidade exponencial. Para $0 < \alpha < 1$, a densidade gama tem a forma da densidade exponencial e a massa da distribuição se concentra próximo de zero. Para $\alpha > 1$, à medida que α aumenta, a massa da distribuição se afasta de zero.

A densidade conjunta de v e u é dada por

$$f(u, v) = \frac{\lambda^\alpha u^{\alpha-1}}{\Gamma(\alpha)\sqrt{2\pi}\sigma_v} \exp\left\{-\lambda u - \frac{v^2}{2\sigma_v^2}\right\}.$$

E a distribuição conjunta de u e ε é

$$f(u,\varepsilon) = \frac{\lambda^{\alpha} u^{\alpha-1}}{\Gamma(\alpha)\sqrt{2\pi}\sigma_v} \exp\left\{-\lambda u - \frac{(\varepsilon+u)^2}{2\sigma^2}\right\}.$$

A função marginal de ε é

$$f(\varepsilon) = \frac{\lambda^{\alpha}\sigma_v^{\alpha-1}}{\Gamma(\alpha)\sqrt{2\pi}} \exp\left\{\lambda\varepsilon + \frac{\lambda^2\sigma_v^2}{2}\right\} \int_w^{\infty} (t-w)^{\alpha-1} \exp\left\{-\frac{t^2}{2}\right\} dt, \tag{2.24}$$

em que $w = (\varepsilon/\sigma_v) + (\lambda\sigma_v)$. A função f_{ε} é assimetricamente distribuída, com média e variância

$$\begin{aligned} \mathrm{E}(\varepsilon) &= -\mathrm{E}(u) = -\frac{\alpha}{\lambda}, \\ \mathrm{V}(\varepsilon) &= \sigma_v^2 + \frac{\alpha}{\lambda^2}. \end{aligned}$$

A equação (2.24) contém um termo integral intratável. O cálculo desse termo é fundamental para o desenvolvimento de $f(\varepsilon)$ e subsequente maximização da função de log-verossimilhança e estimação da ineficiência técnica de cada firma. Por apresentar as dificuldades citadas, o modelo de fronteira estocástica, em que o termo u segue uma distribuição gama, é o menos utilizado na literatura de SFA. Diversos autores dedicaram seus estudos à maximização da função de log-verossimilhança e estimação da eficiência técnica do modelo normal-gama. Este problema não será abordado nesta dissertação. Para um aprofundamento maior neste tema, sugere-se o estudo de Andrade e Souza (2018), que fizeram uma extensa comparação entre seis métodos numéricos de maximização.

Capítulo 3

Método da máxima verossimilhança em dois estágios

3.1 Introdução

Os problemas solucionados pela estimação em dois estágios são encontrados quando elementos de um modelo estão embutidos em outro, por exemplo:

$$\begin{aligned} &\text{Modelo 1}: \quad \{y_1|x_1, \beta_1\} \\ &\text{Modelo 2}: \quad \{y_2|x_2, \beta_2, (y_1|x_1, \beta_1)\}. \end{aligned}$$

Existem dois vetores de parâmetros a serem estimados. O primeiro vetor, β_1, existe nos dois modelos, porém o segundo vetor, β_2 existe apenas no segundo modelo.

Há duas abordagens possíveis para estimar estes vetores. A primeira abordagem é a informação completa de máxima verossimilhança, ICMV, na qual é especificada a distribuição conjunta $f(y_1, y_2|x_1, x_2, \beta_1, \beta_2)$ e maximizada a função de log-verossimilhança conjunta

$$\ln L(\beta_1, \beta_2) = \sum_{i=1}^{n} \ln f(y_{i1}, y_{i2}|x_{i1}, x_{i2}, \beta_1, \beta_2).$$

Alternativamente, pode-se adotar uma informação limitada de máxima verossimilhança, ILMV, procedimento de dois estágios. Nessa abordagem, o primeiro modelo é estimado maximizando

$$\ln L_1(\beta_1) = \sum_{i=1}^{n} \ln f_1(y_{i1}|x_{i1}, \beta_1).$$

Subsequentemente, é estimado o segundo vetor de parâmetros condicionado aos resultados da primeira estimação. Então, a função de log-verossimilhança condicional a ser maximizada é dada por

$$\ln L_2(\hat{\beta}_1, \beta_2) = \sum_{i=1}^{n} \ln f_2 \left\{ y_{2i} | x_{1i}, x_{2i}, \hat{\beta}_1, \beta_2 \right\}. \tag{3.1}$$

Greene (2012) afirma que há pelo menos duas razões para preferir a estimação por máxima verossimilhança em dois estágios. Primeiro, para utilizar a ICMV é necessário especificar a distribuição conjunta, o que pode ser complicado em alguns casos, como, por exemplo, quando um modelo é discreto e o outro contínuo. A segunda razão é que maximizar a log-verossimilhança conjunta costuma ser numericamente e computacionalmente difícil. Maximizar as log-verossimilhanças separadas pode ser mais simples e direto.

Entretanto, a estimação em dois estágios também carrega alguns problemas. Murphy e Topel (2002) afirmam que procedimentos em dois estágios produzem estimativas consistentes dos parâmetros do modelo do segundo estágio sob condições gerais. Porém, a estimativa dos erros padrão e estatísticas de testes relacionados a eles são incorretas.

O argumento para consistência de $\hat{\beta}_2$ é essencialmente que, se β_1 fosse conhecido, todos os resultados de estimação por máxima verossimilhança poderiam ser aplicados a estimação de β_2, e também porque, assintoticamente, $\hat{\beta}_1 \xrightarrow{p} \beta_1$. Porém, o mesmo raciocínio não pode ser aplicado para justificar o uso de $(1/n)\hat{\mathrm{V}}_2$ como estimador da matriz assintótica de covariância de $\hat{\beta}_2$. Dessa forma, é necessária uma correção que considere que β_1 está sendo utilizado para estimação de β_2. Greene (2012) descreve os resultados de Murphy e Topel (2002) a esse respeito, o que será detalhado na próxima seção.

3.2 *Estimador de Murphy-Topel para variância em um modelo de dois estágios*

Se as condições de regularidade são satisfeitas para ambas as funções de log-verossimilhança, então o estimador de máxima verossimilhança do segundo estágio para β_2 é consistente e assintoticamente normalmente distribuído com matriz de variância-covariância dada por

$$\mathrm{V}_2^* = \frac{1}{n}\left[\mathrm{V}_2 + \mathrm{V}_2\left[\mathrm{C}\mathrm{V}_1\mathrm{C}' - \mathrm{R}\mathrm{V}_1\mathrm{C}' - \mathrm{C}\mathrm{V}_1\mathrm{R}'\right]\mathrm{V}_2\right], \tag{3.2}$$

em que

$$\mathrm{V}_1 = \text{Variância assintótica de } \hat{\beta}_1 \text{ baseada em } \ln L_1,$$

$$\mathrm{V}_2 = \text{Variância assintótica de } \hat{\beta}_2 \text{ baseada em } \ln L_2|\beta_1,$$

$$\mathrm{C} = \mathrm{E}\left[\frac{1}{n}\left(\frac{\partial \ln L_2}{\partial \beta_2}\right)\left(\frac{\partial \ln L_2}{\partial \beta_1'}\right)\right],$$

$$\mathrm{R} = \mathrm{E}\left[\frac{1}{n}\left(\frac{\partial \ln L_2}{\partial \beta_2}\right)\left(\frac{\partial \ln L_1}{\partial \beta_1'}\right)\right].$$

A correção da matriz de covariância assintótica no segundo estágio requer uma computação adicional. As matrizes V_1 e V_2 podem ser estimadas de duas formas. Hardin (2002) estimou $\hat{\mathrm{V}}_1$ e $\hat{\mathrm{V}}_2$ conforme as equações (3.3) e (3.4). Esta é a maneira mais usual para modelos estimados por máxima verossimilhança. Na programação desenvolvida no R para aplicação desta dissertação, optou-se por essa abordagem.

$$\hat{\mathrm{V}}_1 = \left[-\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial^2 \ln f_{i1}}{\partial \hat{\beta}_1 \hat{\beta}_1'}\right)\right]^{-1} \tag{3.3}$$

e

$$\hat{\mathrm{V}}_2 = \left[-\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial^2 \ln f_{i2}}{\partial \hat{\beta}_2 \hat{\beta}_2'}\right)\right]^{-1}. \tag{3.4}$$

Em outros contextos econométricos, faz-se uso dos estimadores BHHH (calculados por produto externo de gradientes) para estimar V_1 e V_2, conforme as equações (3.5) e (3.6) apresentadas em Greene (1990)

$$\hat{\mathrm{V}}_1 = \left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{i1}}{\partial \hat{\beta}_1}\right)\left(\frac{\partial \ln f_{i1}}{\partial \hat{\beta}_1'}\right)\right]^{-1} \tag{3.5}$$

e

$$\hat{\mathrm{V}}_2 = \left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{i2}}{\partial \hat{\beta}_2}\right)\left(\frac{\partial \ln f_{i2}}{\partial \hat{\beta}_2'}\right)\right]^{-1}. \tag{3.6}$$

As matrizes R e C são obtidas somando as observações individuais do produto cruzado das derivadas. Elas são estimadas por

$$\hat{\mathrm{C}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{i2}}{\partial \hat{\beta}_2}\right)\left(\frac{\partial \ln f_{i2}}{\partial \hat{\beta}_1'}\right)$$

e

$$\hat{\mathrm{R}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{i2}}{\partial \hat{\beta}_2}\right)\left(\frac{\partial \ln f_{i1}}{\partial \hat{\beta}_1'}\right).$$

3.3 Demonstração

Greene (1990) apresenta a demonstração desses resultados. O estimador de máxima verossimilhança do primeiro estágio é definido por

$$\frac{1}{n}\frac{\partial \ln L_1(\hat{\beta}_1)}{\partial \hat{\beta}_1} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \ln f_1(y_{i1}|x_{i1},\hat{\beta}_1)}{\partial \hat{\beta}_1} = \frac{1}{n}\sum_{i=1}^{n} g_{i1}(\hat{\beta}_1) = \bar{g}_1(\hat{\beta}_1) = 0.$$

Usando os resultados do apêndice E, a sua distribuição assintótica é

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} \left[-\mathrm{H}_{11}^{(1)}(\beta_1)\right]^{-1}\sqrt{n}\bar{g}_1(\beta_1),$$

em que a expressão significa que a distribuição limite de dois vetores aleatórios é a mesma, e

$$\mathrm{H}_{11}^{(1)} = \mathrm{E}\left[\frac{1}{n}\frac{\partial^2 \ln L_1(\beta_1)}{\partial \beta_1 \partial \beta_1'}\right].$$

O estimador de máxima verossimilhança do segundo estágio de β_2 é definido por

$$\frac{1}{n}\frac{\partial \ln L_2(\hat{\beta}_1,\hat{\beta}_2)}{\partial \hat{\beta}_2} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \ln f_2(y_{i2}|x_{i1},x_{i2},\hat{\beta}_1,\hat{\beta}_2)}{\partial \hat{\beta}_2} = \frac{1}{n}\sum_{i=1}^{n} g_{i2}(\hat{\beta}_1,\hat{\beta}_2) = \bar{g}_2(\hat{\beta}_1,\hat{\beta}_2) = 0.$$

Ao expandir o vetor de derivadas, $\bar{g}_2(\hat{\beta}_1,\hat{\beta}_2)$, em uma série de Taylor linear, e usar novamente os resultados do apêndice E, obtém-se

$$\bar{g}_2(\hat{\beta}_1,\hat{\beta}_2) = \bar{g}_2(\beta_1,\beta_2) + \left[\mathrm{H}_{22}^{(2)}(\beta_1,\beta_2)\right](\hat{\beta}_2 - \beta_2) + \left[\mathrm{H}_{21}^{(2)}(\beta_1,\beta_2)\right](\hat{\beta}_1 - \beta_1) + o(1/n) = 0,$$

em que

$$\mathrm{H}_{21}^{(2)}(\beta_1,\beta_2) = \mathrm{E}\left[\frac{1}{n}\frac{\partial^2 \ln L_2(\beta_1,\beta_2)}{\partial \beta_2 \partial \beta_1'}\right]$$

e

$$\mathrm{H}_{22}^{(2)}(\beta_1,\beta_2) = \mathrm{E}\left[\frac{1}{n}\frac{\partial^2 \ln L_2(\beta_1,\beta_2)}{\partial \beta_2 \partial \beta_2'}\right].$$

A distribuição assintótica é dada por

$$\sqrt{n}(\hat{\beta}_2 - \beta_2) \xrightarrow{d} \left[-\mathrm{H}_{22}^{(2)}(\beta_1,\beta_2)\right]^{-1}\sqrt{n}\bar{g}_2(\beta_1,\beta_2) + \left[-\mathrm{H}_{22}^{(2)}(\beta_1,\beta_2)\right]^{-1}\left[\mathrm{H}_{21}^{(2)}(\beta_1,\beta_2)\right]\sqrt{n}(\hat{\beta}_1 - \beta_1).$$

Por conveniência, denote $\mathrm{H}_{22}^{(2)} = \mathrm{H}_{22}^{(2)}(\beta_1,\beta_2)$, $\mathrm{H}_{21}^{(2)} = \mathrm{H}_{21}^{(2)}(\beta_1,\beta_2)$ e $\mathrm{H}_{11}^{(1)} = \mathrm{H}_{11}^{(1)}(\beta_1)$. Agora, substitua o estimador do primeiro estágio para β_1 nesta expressão:

$$\sqrt{n}(\hat{\beta}_2 - \beta_2) \xrightarrow{d} \left[-\mathrm{H}_{22}^{(2)}\right]^{-1}\sqrt{n}\bar{g}_2(\beta_1,\beta_2) + \left[-\mathrm{H}_{22}^{(2)}\right]^{-1}\left[\mathrm{H}_{21}^{(2)}\right]\left[-\mathrm{H}_{11}^{(1)}\right]^{-1}\sqrt{n}\bar{g}_1(\beta_1).$$

Para obter a matriz de covariância de $\hat{\beta}_2$ é necessário obter a variância limite do vetor aleatório na expressão precedente. A distribuição normal conjunta dos dois primeiros vetores de derivadas tem média zero e

$$\text{Var} = \begin{bmatrix} \sqrt{n}\bar{g}_1(\beta_1) \\ \sqrt{n}\bar{g}_2(\beta_1, \beta_2) \end{bmatrix} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

Então, a matriz de covariância desejada é

$$\begin{aligned} \text{Var}[\sqrt{n}(\hat{\beta}_2 - \beta_2)] &= \left[-\text{H}_{22}^{(2)}\right]^{-1} \Sigma_{22} \left[-\text{H}_{22}^{(2)}\right]^{-1} \\ &+ \left[-\text{H}_{22}^{(2)}\right]^{-1} \left[\text{H}_{21}^{(2)}\right] \left[-\text{H}_{11}^{(1)}\right]^{-1} \Sigma_{11} \left[-\text{H}_{11}^{(1)}\right]^{-1} \left[\text{H}_{21}^{(2)}\right]' \left[-\text{H}_{22}^{(2)}\right]^{-1} \\ &+ \left[-\text{H}_{22}^{(2)}\right]^{-1} \Sigma_{21} \left[-\text{H}_{11}^{(1)}\right]^{-1} \left[\text{H}_{21}^{(2)}\right]' \left[-\text{H}_{22}^{(2)}\right]^{-1} \\ &+ \left[-\text{H}_{22}^{(2)}\right]^{-1} \left[\text{H}_{21}^{(2)}\right] \left[-\text{H}_{11}^{(1)}\right]^{-1} \Sigma_{12} \left[-\text{H}_{22}^{(2)}\right]^{-1}. \end{aligned}$$

A variância do vetor de derivadas primeiras de log-verossimilhança é a matriz negativa da esperança de derivadas segundas (Veja o apêndice D). Logo, $\Sigma_{22} = \left[-\text{H}_{22}^{(2)}\right]$ e $\Sigma_1 = \left[-\text{H}_{11}^{(1)}\right]$. Substituindo, obtém-se

$$\begin{aligned} \text{Var}[\sqrt{n}(\hat{\beta}_2 - \beta_2)] &= \left[-\text{H}_{22}^{(2)}\right]^{-1} + \left[-\text{H}_{22}^{(2)}\right]^{-1} \left[\text{H}_{21}^{(2)}\right] \left[-\text{H}_{11}^{(1)}\right]^{-1} \left[\text{H}_{21}^{(2)}\right]' \left[-\text{H}_{22}^{(2)}\right]^{-1} \\ &+ \left[-\text{H}_{22}^{(2)}\right]^{-1} \Sigma_{21} \left[-\text{H}_{11}^{(1)}\right]^{-1} \left[\text{H}_{21}^{(2)}\right]' \left[-\text{H}_{22}^{(2)}\right]^{-1} \\ &+ \left[-\text{H}_{22}^{(2)}\right]^{-1} \left[\text{H}_{21}^{(2)}\right] \left[-\text{H}_{11}^{(1)}\right]^{-1} \Sigma_{12} \left[-\text{H}_{22}^{(2)}\right]^{-1}. \end{aligned}$$

De (E.1), $\left[-\text{H}_{11}^{(1)}\right]^{-1}$ *e* $\left[-\text{H}_{22}^{(2)}\right]^{-1}$ são V_1 e V_2 da equação (3.2), o que reduz a expressão para

$$\begin{aligned} \text{Var}[\sqrt{n}(\hat{\beta}_2 - \beta_2)] &= \text{V}_2 + \text{V}_2 \left[\text{H}_{21}^{(2)}\right] \text{V}_1 \left[\text{H}_{21}^{(2)}\right]' \text{V}_2 \\ &- \text{V}_2 \Sigma_{21} \text{V}_1 \left[\text{H}_{21}^{(2)}\right]' \text{V}_2 \\ &- \text{V}_2 \left[\text{H}_{21}^{(2)}\right] \text{V}_1 \Sigma_{12} \text{V}_2. \end{aligned}$$

Os dois termos remanescentes são $\text{H}_{21}^{(2)}$, que é a $\text{E}\left[\frac{1}{n}\frac{\partial^2 \ln L_2(\beta_1,\beta_2)}{\partial\beta_2\partial\beta_1'}\right]$, estimado por $-\text{C}$ em (3.2). E Σ_{21}, que é a covariância dos dois vetores de primeiras derivadas, estimado por R, também em (3.2). Então, completando a demonstração

$$\text{Var}[\sqrt{n}(\hat{\beta}_2 - \beta_2)] = \text{V}_2 + \text{V}_2\text{C}\text{V}_1\text{C}'\text{V}_2 - \text{V}_2\text{R}\text{V}_1\text{C}'\text{V}_2 - \text{V}_2\text{C}\text{V}_1\text{R}'\text{V}_2.$$

3.4 Log-Verossimilhanças

Para exemplificar as log-verossimilhanças que são calculadas para correção de Murphy Topel, considere uma estimação em dois estágios em que o primeiro estágio assume uma regressão logística $y_{i1} = \mathrm{E}\{y_{i1}\} + \varepsilon_{i1}$, em que

$$\mathrm{E}\{y_{i1}\} = \pi_i = \frac{\exp(x_{i1}^{'}\beta_1)}{1+\exp(x_{i1}^{'}\beta_1)} = \frac{1}{1+\exp(-x_{i1}^{'}\beta_1)}.$$

Seja $\eta = x_{i1}^{'}\beta_1$, $P(y_{i1}=1) = \pi_i$, $P(y_{i1}=0) = 1-\pi_i$ e $f_{i1}(y_{i1}) = \pi_i^{y_{i1}}(1-\pi_i)^{1-y_{i1}}$. A densidade conjunta de n observações é dada por

$$g(y_1, ..., y_n) = \prod_{i=1}^{n} \pi_i^{y_{i1}}(1-\pi_i)^{1-y_{i1}}$$

e a função de log-verossimilhança será dada por

$$L_1 = \sum_{i=1}^{n} \left(y_{i1}\log(\pi_i) + (1-y_{i1})\log(1-\pi_i)\right). \tag{3.7}$$

Agora considere que o segundo estágio é um modelo de fronteira estocástica de produção, em que o componente de ineficiência contém a probabilidade estimada no primeiro estágio, $\hat{\pi}_i = \frac{1}{1+\exp(-x_{i1}^{'}\hat{\beta}_1)}$. Então,

$$y_{i2} = x_{i2}^{'}\beta_2 + \varepsilon_{i2}$$

em que, $\varepsilon_{i2} = v_i - u_i = y_{i2} - x_{i2}^{'}\beta_2$, $v_i \sim N(0, \sigma_v^2)$, $u_i \sim N_+(0, \sigma_u^2)$. Considere ainda que $\sigma_u^2 = \exp(\alpha)$, $\alpha = m_i^{'}\delta$, em que $m_i^{'}$ é um vetor que contém as variáveis, inclusive a probabilidade estimada no primeiro estágio. Utilizando (2.12), a função de log-verossimilhança do segundo estágio pode ser reescrita como

$$\begin{aligned} L_2 = \; & - \frac{1}{2}n\log\left(\frac{\pi}{2}\right) - \frac{1}{2}n\log(\sigma_v^2+\sigma_u^2) \\ & + \sum_{i=1}^{n}\log\Phi\left(-\frac{\sigma_u\epsilon_i}{\sigma_v\sqrt{\sigma_v^2+\sigma_u^2}}\right) - \frac{1}{2\sigma_v^2+2\sigma_u^2}\sum_{i=1}^{n}\epsilon_i^2. \end{aligned} \tag{3.8}$$

3.5 Gradientes

O próximo passo, para calcular a correção de Murphy-Topel, é calcular as derivadas das funções de verossimilhança dos primeiro e segundo estágios, que compõe as matrizes R e C.

Para o primeiro estágio temos:

$$\begin{aligned}
\frac{\partial L_1}{\partial \beta_1} &= \sum_{i=1}^{n}(y_i - \hat{\pi}_i)x_{i1}^{'}, \\
\frac{\partial L_1}{\partial \beta_2} &= 0, \\
\frac{\partial L_1}{\partial \sigma_v^2} &= 0 \text{ e} \\
\frac{\partial L_1}{\partial \sigma_u^2} &= 0.
\end{aligned}$$

Para o segundo estágio temos:

$$\begin{aligned}
\frac{\partial L_2}{\partial \beta_2} &= \sum_{i=1}^{n} x_{i2}^{'}\left(\frac{\varepsilon_i}{\sigma^2} + A_i\frac{\lambda}{\sigma}\right), \\
\frac{\partial L_2}{\partial \sigma_u^2} &= \sum_{i=1}^{n}\frac{1}{2\sigma^2}\left(\frac{\varepsilon_i^2}{\sigma^2} - A_i\frac{\varepsilon_i}{\lambda\sigma} - 1\right), \\
\frac{\partial L_2}{\partial \sigma_v^2} &= \sum_{i=1}^{n}\frac{1}{2\sigma^2}\left(\frac{\varepsilon_i^2}{\sigma^2} + A_i\frac{\varepsilon_i\lambda}{\sigma}(2+\lambda^2) - 1\right), \\
\frac{\partial L_2}{\partial \beta_1} &= \sum_{i=1}^{n}\frac{1}{2\sigma^2}\left(\frac{\varepsilon_i^2}{\sigma^2} - A_i\frac{\varepsilon_i}{\lambda\sigma} - 1\right)\exp(\alpha)\hat{\delta}_{\hat{\theta}_1}\hat{\pi}_i(1-\hat{\pi}_i)m_i^{'},
\end{aligned}$$

em que, $\lambda = \frac{\sigma_u}{\sigma_v}$, $\hat{\delta}_{\hat{\beta}_1}$ é o coeficiente estimado do modelo do segundo estágio para o preditor estimado no primeiro estágio $\hat{\pi}_i$, $A_i = \frac{\phi(a)}{\Phi(a)}$ e $a = \frac{\varepsilon_i\lambda}{\sigma}$.

Capítulo 4

Base de dados

A teoria de fronteira estocástica de produção tem sido aplicada em várias áreas além da Teoria de Firma. Uma delas é a área educacional, como observaram Rosa et al. (2016). Neste trabalho, Rosa et al. (2016) realizaram um modelo de fronteira estocástica de produção para avaliar a eficiência técnica de escolas brasileiras. Com o objetivo de produzir análise semelhante, porém a partir da estimação em dois estágios, a base de dados construída para a aplicação desta dissertação reuniu microdados do Censo Escolar, da Prova Brasil, que compõe o Sistema de Avaliação da Educação Básica (Saeb), além de indicadores disponibilizados no sítio eletrônico do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep).

As variáveis utilizadas e tratamentos necessários para compor a base de dados serão relatados nas próximas seções.

4.1 Microdados

4.1.1 Censo Escolar

O Censo Escolar é o principal instrumento de coleta de informações da educação básica e a mais importante pesquisa estatística educacional brasileira. É coordenado pelo Inep e realizado em regime de colaboração entre as secretarias estaduais e municipais de educação e com a participação de todas as escolas públicas e privadas do país.

Ele abrange as diferentes etapas e modalidades da educação básica e profissional: ensino regular (educação infantil, ensino fundamental e médio); educação especial – modalidade substitutiva; educação de jovens e adultos (EJA); educação profissional (cursos técnicos e cursos de formação inicial continuada ou qualificação profissional).

O Censo Escolar é uma ferramenta fundamental para que os atores educacionais possam compreender a situação educacional do país, das unidades federativas, dos municípios e do Distrito Federal, bem como das escolas e, com isso, acompanhar a efetividade das políticas públicas.

4.1.1.1 Variáveis e tratamentos

Para este trabalho, foram selecionadas as escolas públicas do Censo Escolar de 2015 que ofertaram o 5º ano do ensino fundamental.

A partir dos dados do Censo Escolar, foram utilizadas as seguintes informações: número de docentes por aluno, número de computadores por aluno, existência de biblioteca e/ou sala de leitura na escola, tempo médio de duração das turmas em minutos, localização da escola (urbana ou rural), dependência administrativa da escola e código da UF onde está localizada a escola.

4.1.2 Sistema Nacional de Avaliação do Rendimento Escolar - Prova Brasil

O Saeb é composto por três avaliações externas em larga escala que permitem ao Inep realizar um diagnóstico da educação básica brasileira e de alguns fatores que possam interferir no desempenho do estudante, fornecendo um indicativo sobre a qualidade do ensino ofertado.

As três avaliações são: Avaliação Nacional do Rendimento Escolar (Anresc), conhecida como Prova Brasil, criada com o objetivo de avaliar a qualidade do ensino ministrado nas escolas das redes públicas. Avaliação Nacional da Educação Básica (Aneb) que utiliza os mesmos instrumentos da Prova Brasil, é aplicada com a mesma periodicidade, mas diferencia-se por abranger, de forma amostral, escolas e alunos das redes públicas e privadas do País que não atendem aos critérios de participação da Prova Brasil. E Avaliação Nacional da Alfabetização (ANA) que tem como objetivo aferir os níveis de alfabetização e letramento em Língua Portuguesa (leitura e escrita) e Matemática.

Nesta dissertação, serão utilizados os dados da Prova Brasil de 2015. A Prova Brasil é uma avaliação censitária bianual envolvendo os alunos do 5º ano (4ª série) e 9º ano (8ª série) do Ensino Fundamental das escolas públicas que possuem, no mínimo, 20 alunos matriculados nas séries/anos avaliados. Seu objetivo principal é mensurar a qualidade do ensino ministrado nas escolas das redes públicas. Produz informações sobre os níveis de

aprendizagem em Língua Portuguesa (Leitura) e em Matemática e fornece resultados para cada unidade escolar participante bem como para as redes de ensino em geral. Apresenta, ainda, indicadores contextuais sobre as condições extra e intraescolares. Os dados apresentados visam servir de subsídio para diagnóstico, reflexão e planejamento do trabalho pedagógico da escola, bem como para a formulação de ações e políticas públicas com vistas à melhoria da qualidade da educação básica.

4.1.2.1 Variáveis e tratamentos

O Inep calcula e disponibiliza as notas médias aferidas pelas escolas em português e matemática a partir das notas obtidas por seus alunos em cada uma destas disciplinas. Para esta dissertação, usando como exemplo o estudo feito pela Codeplan (Rosa et al. (2016)), foi calculada a nota média da escola a partir das notas médias de português e matemática.

Considerando a metodologia utilizada por Rosa et al. (2016), foram calculados quatro índices a partir das respostas dadas aos questionários contextuais da Prova Brasil, para compor a base de dados . São eles: índice de histórico escolar do aluno, índice de atenção dos pais, índice de dedicação extra classe, índice de acesso à cultura. Os índices utilizados foram construídos a partir de um conjunto de variáveis categóricas que qualificavam as condições dos itens considerados. Foi utilizada uma análise por componentes principais para construir cada um deles, considerando a matriz de correlações policóricas. A correlação policórica é a generalização da correlação tetracórica para uma matriz n x m. Holgado-Tello et al. (2010) discorreram sobre esta correlação, conforme apresentado no apêndice F. Já a correlação tetracórica é a correlação de Pearson inferida de uma tabela 2 x 2 com a suposição de normalidade bivariada. As variáveis que compõe os índices calculados estão listadas no apêndice A.1.

4.1.3 Indicadores

Além dos microdados do Censo Escolar e da Prova Brasil, o Inep também disponibiliza os microdados de indicadores elaborados a partir de dados destes instrumentos.

Para compor a base de dados, foram selecionados cinco indicadores: indicador de adequação da formação do docente (AFD), indicador de esforço docente (IED), indicador de regularidade docente (IRD), indicador de complexidade de gestão da escola (ICG) e índice

de nível socioeconômico (Inse).

4.1.3.1 Indicador de adequação da formação do docente

A partir dos dados do Censo Escolar, o Inep desenvolveu o indicador de adequação da formação do docente da educação básica (Inep (2014a)). Para cada uma das disciplinas analisadas foi identificada a formação do docente responsável por seu desenvolvimento na turma, o que possibilitou a formação de cinco grupos de regência das disciplinas. Para este indicador, o Inep disponibiliza o percentual de docentes em cada grupo por escola.

Para fins desta dissertação, foi utilizada como variável a soma dos percentuais de docentes nos grupos um e dois do indicador por unidade escolar.

Tabela 4.1 - Indicador de adequação da formação do docente

Grupo	Descrição
1	Docentes com formação superior de licenciatura na mesma disciplina que lecionam, ou bacharelado na mesma disciplina com curso de complementação pedagógica concluído.
2	Docentes com formação superior de bacharelado na disciplina correspondente, mas sem licenciatura ou complementação pedagógica.
3	Docentes com licenciatura em área diferente daquela que leciona, ou com bacharelado nas disciplinas da base curricular comum e complementação pedagógica concluída em área diferente daquela que leciona.
4	Docentes com outra formação superior não considerada nas categorias anteriores.
5	Docentes que não possuem curso superior completo.

4.1.3.2 Indicador de esforço docente

O indicador de esforço docente também foi construído a partir de dados do Censo Escolar (Inep (2014b)). Este indicador busca sintetizar aspectos do trabalho do professor que contribuem para a sobrecarga no exercício da profissão. Foram utilizadas as informações de números de turnos de trabalho, escolas e etapas de atuação, além da quantidade de alunos atendidos na Educação Básica. O Inep disponibiliza por escola o percentual de docentes em cada nível do indicador.

Assim como no indicador de adequação da formação, para esta dissertação, foi utilizada como variável a soma dos percentuais de docentes nos níveis um e dois do indicador por unidade escolar.

Tabela 4.2 - Indicador de esforço docente

Níveis	Descrição
Nível 1	Docente que tem até 25 alunos e atua em um único turno, escola e etapa.
Nível 2	Docente que tem entre 25 e 150 alunos e atua em um único turno, escola e etapa.
Nível 3	Docente que tem entre 25 e 300 alunos e atua em um ou dois turnos em uma única escola e etapa.
Nível 4	Docentes que tem entre 50 e 400 alunos e atua em dois turnos, em uma ou duas escolas e em duas etapas.
Nível 5	Docente que tem mais de 300 alunos e atua nos três turnos, em duas ou três escolas e em duas ou três etapas.
Nível 6	Docente que tem mais de 400 alunos e atua nos três turnos, em duas ou três escolas e em duas etapas ou três etapas.

4.1.3.3 Indicador de regularidade docente

Neste indicador, também construído a partir de dados do Censo Escolar, para cada par professor-escola é atribuída uma pontuação de forma que sejam consideradas a presença do docente em anos mais recentes e a regularidade em anos consecutivos (Inep (2015)). O IRD é definido como a pontuação final de cada par professor-escola, padronizada para variar em uma escala de 0 a 5. Assim, quanto mais próximo de 0, mais irregular é o professor, e quanto mais próximo de 5, mais regular é o professor. O Inep disponibiliza, por escola, a pontuação média dos pares professor-escola.

4.1.3.4 Indicador de complexidade de gestão da escola

Neste indicador, que também utiliza os dados do Censo Escolar, assume-se que a complexidade da gestão escolar se concretiza em quatro características (Inep (2014d)): (1) porte da escola; (2) número de turnos de funcionamento; (3) complexidade das etapas ofertadas pela escola e (4) número de etapas/modalidades oferecidas. Desta forma, o Inep categorizou a complexidade de gestão da escola em seis níveis.

Para fins desta dissertação, uma nova variável com três categorias foi gerada, agregando na primeira categoria os níveis um e dois do ICG, na segunda categoria os níveis três e quatro, e na terceira categoria os níveis cinco e seis.

Tabela 4.3 - Indicador de complexidade de gestão da escola

Níveis	Descrição
Nível 1	Porte inferior a 50 matrículas, operando em único turno e etapa e apresentando a Educação Infantil ou Anos Iniciais como etapa mais elevada.
Nível 2	Porte entre 50 e 300 matrículas, operando em 2 turnos, com oferta de até 2 etapas e apresentando a Educação Infantil ou Anos Iniciais como etapa mais elevada.
Nível 3	Porte entre 50 e 500 matrículas, operando em 2 turnos, com 2 ou 3 etapas e apresentando os Anos Finais como etapa mais elevada.
Nível 4	Porte entre 150 e 1000 matrículas, operando em 2 ou 3 turnos, com 2 ou 3 etapas, apresentando o Ensino Médio/profissional ou a EJA como etapa mais elevada.
Nível 5	Porte entre 150 e 1000 matrículas, operando em 3 turnos, com 2 ou 3 etapas, apresentando a EJA como etapa mais elevada.
Nível 6	Porte superior a 500 matrículas, operando em 3 turnos, com 4 ou mais etapas, apresentando a EJA como etapa mais elevada.

4.1.3.5 Indicador de nível socioeconômico das escolas

Este indicador foi construído a partir dos dados do Saeb (Inep (2014c)), a partir das respostas dadas pelos alunos aos questionários contextuais da Aneb, da Prova Brasil e do Exame Nacional do Ensino Médio (Enem), referentes aos anos de 2011 e 2013. O universo de referência do Inse inclui somente os dados dos estudantes dessas bases que responderam, ao preencher o questionário contextual, cinco ou mais questões, referentes a:

- posse de bens no domicílio: televisão em cores, tv por assinatura, telefone fixo, telefone celular, acesso a internet, aspirador de pó, rádio, videocassete ou DVD, geladeira, freezer (aparelho independente ou parte da geladeira duplex), máquina de lavar roupa, carro, computador, quantidade de banheiros e quartos para dormir;
- contratação de serviços: contratação de serviços de mensalista ou diarista;

- renda: renda familiar mensal, em salários mínimos;
- escolaridade: escolaridade do pai e escolaridade da mãe.

O nível socioeconômico da escola foi definido como a média aritmética simples da medida de nível socioeconômico de seus respectivos alunos e, em seguida, para melhor representar os conjuntos de escolas com mais de 10 alunos na base de dados, foram criados, a partir da análise de cluster (K-means), sete níveis, classificados da seguinte maneira: muito baixo, médio baixo, baixo, médio, médio alto, alto e muito alto.

Para fins desta dissertação, uma nova variável foi criada, com apenas duas categorias. Na primeira, agrega-se os níveis muito baixo, médio baixo e baixo e na segunda categoria agrega-se os demais níveis.

Destaca-se que esse foi o único indicador cujo ano de referência da divulgação não é 2015, porém, acredita-se que o contexto socioeconômico médio dos alunos de determinada escola não sofra alterações significativas nesse espaço de tempo.

Tabela 4.4 - Indicador de nível socioeconômico

Níveis	Descrição
Muito baixo	Este é o menor nível da escala e os alunos, de modo geral, indicaram que há em sua casa bens elementares, como uma televisão em cores, uma geladeira, um telefone celular, até dois quartos no domicílio e um banheiro; não contratam empregada mensalista e nem diarista; a renda familiar mensal é de até 1 salário mínimo; e seus pais ou responsáveis possuem ensino fundamental completo ou estão cursando cursando esse nível de ensino.
Médio baixo	Neste, os alunos, de modo geral, indicaram que há em sua casa bens elementares, como uma televisão em cores, um rádio, uma geladeira, um telefone celular, dois quartos e um banheiro; bem complementar, como videocassete ou DVD; não contratam empregada mensalista e nem diarista; a renda familiar mensal é de até 1 salário mínimo; e seus pais ou responsáveis possuem ensino fundamental completo ou estão cursando esse nível de ensino.

Tabela 4.4 - Indicador de nível socioeconômico

Níveis	Descrição
Baixo	Neste, os alunos, de modo geral, indicaram que há em sua casa bens elementares, como uma televisão em cores, um rádio, uma geladeira, um telefone celular, dois quartos e um banheiro; bens complementares, como videocassete ou DVD, máquina de lavar roupas, computador e possuem acesso à internet; não contratam empregada mensalista ou diarista; a renda familiar mensal está entre 1 e 1,5 salários mínimos; e seu pai e sua mãe (ou responsáveis) possuem ensino fundamental completo ou estão cursando esse nível de ensino.
Médio	Já neste nível, os alunos, de modo geral, indicaram que há em sua casa bens elementares, como um rádio, uma geladeira, dois telefones celulares, até dois quartos e um banheiro e, agora, duas ou mais televisões em cores; bens complementares, como videocassete ou DVD, máquina de lavar roupas, computador e possuem acesso à internet; bens suplementares, como freezer, um ou mais telefones fixos e um carro; não contratam empregada mensalista ou diarista; a renda familiar mensal está entre 1,5 e 5 salários mínimos; e seu pai e sua mãe (ou responsáveis) possuem ensino fundamental completo ou estão cursando esse nível de ensino.
Médio alto	Neste, os alunos, de modo geral, indicaram que há em sua casa um quantitativo maior de bens elementares como três quartos e dois banheiros; bens complementares, como videocassete ou DVD, máquina de lavar roupas, computador e acesso à internet; bens suplementares, como freezer, um ou mais telefones fixos, um carro, além de uma TV por assinatura e um aspirador de pó; não contratam empregada mensalista ou diarista; a renda familiar mensal é maior, pois está entre 5 e 7 salários mínimos; e seu pai e sua mãe (ou responsáveis) completaram o ensino médio.

Tabela 4.4 - Indicador de nível socioeconômico

Níveis	Descrição
Alto	Neste nível, os alunos, de modo geral, indicaram que há em sua casa um quantitativo alto de bens elementares como três quartos e três banheiros; bens complementares, como videocassete ou DVD, máquina de lavar roupas, computador e acesso à internet; bens suplementares, como freezer, telefones fixos, uma TV por assinatura, um aspirador de pó e, agora, dois carros; não contratam empregada mensalista ou diarista; a renda familiar está acima de 7 salários mínimos; e seu pai e sua mãe (ou responsáveis) completaram a faculdade e/ou podem ter concluído ou não um curso de pós-graduação.
Muito alto	Este é o maior nível da escala e os alunos, de modo geral, indicaram que há em sua casa um quantitativo alto de bens elementares, como duas ou mais geladeiras e três ou mais televisões em cores, por exemplo; bens complementares, como videocassete ou DVD, máquina de lavar roupas, computador e acesso à internet; maior quantidade de bens suplementares, tal como três ou mais carros e TV por assinatura; contratam, também, empregada mensalista ou diarista até duas vezes por semana; a renda familiar mensal é alta, pois está acima de 7 salários mínimos; e seu pai e sua mãe (ou responsáveis) completaram a faculdade e/ou podem ter concluído ou não um curso de pós-graduação.

Capítulo 5

Aplicação

Assim como no estudo de Rosa et al. (2016), definiu-se como produto para o modelo de fronteira estocástica a média das notas das disciplinas português e matemática da Prova Brasil de 2015 para o 5° ano do ensino fundamental.

Como insumos foram selecionadas as variáveis: número de docentes por aluno, número de computadores por alunos, tempo médio de duração das turmas, existência de sala de leitura e/ou biblioteca na escola. Essas são variáveis já coletadas pelo Censo Escolar e entende-se que são insumos diretos na função de produção das escolas.

Para modelar a ineficiência foram selecionadas variáveis que podem impactar a média da nota da Prova Brasil, mas sobre as quais as escolas não detêm o controle. São elas: índice de histórico escolar dos alunos, índice de atenção dos pais, índice de acesso à cultura, índice de dedicação extraclasse, indicador de formação docente, indicador de esforço docente, indicador de complexidade de gestão da escola, indicador de regularidade docente, indicador do nível socioeconômico das escolas. Em comparação ao estudo de Rosa et al. (2016), foi possível, a partir dos quatro indicadores lançados pelo Inep, reduzir o número de variáveis que tratam sobre a formação e esforço de trabalho do docente, e incluir ainda uma variável que captasse a complexidade de gestão da escola.

Por fim, optou-se por controlar a variância do termo de erro aleatório bilateral por UF e por localidade, uma vez que essas são características externas que podem produzir possíveis heterogeneidades.

A implementação da análise foi desenvolvida no software RStudio, versão 1.1.456. Apesar do RStudio ter alguns pacotes que podem ser utilizados para o cálculo de fronteiras estocásticas, não há rotina pública disponível que aborde a estimação em dois estágios e a correção de Murphy-Topel simultaneamente. Com isso, foi desenvolvida uma programação

que contemplasse essas técnicas.

5.1 Análise descritiva

Na base de dados utilizada, há informação da nota média da Prova Brasil para 38.388 escolas públicas brasileiras (federais, estaduais e municipais) com alunos no 5° ano. Após agregar todas as variáveis na base de dados, há uma perda de cerca de 6,6%, pois nem todas as escolas têm informação preenchida para todas as variáveis, resultando em um total de 35.853 escolas no estudo.

Nas tabelas (5.1), (5.2), (5.3), (5.4) e (5.5) são apresentados os números absolutos das variáveis categóricas. Verifica-se uma prevalência de escolas com nível socioeconômico médio a muito alto, com biblioteca e na área urbana. Em relação à complexidade de gestão, o número de escolas nos níveis 1 e 2, 5 e 6 é semelhante, enquanto nos níveis 3 e 4 é um pouco maior. Em relação à distribuição de escolas por UF, estados mais populosos, como Bahia, Minas Gerais e São Paulo têm mais escolas, como esperado.

Em relação aos boxplots apresentados, é possível observar que quanto maior o nível socioeconômico dos alunos da escola, maior é também a nota média da Prova Brasil (5.1). O boxplot da média da nota da Prova Brasil pelo ICG indica que escolas com gestão mais complexa apresentam piores resultados no teste de proficiência (5.2). E a figura (5.4) indica que escolas com biblioteca e/ou sala de leitura tem melhor desempenho. Já as figuras (5.3) e (5.5) corroboram com a tese de que o estado que a escola pertence e a localização urbana ou rural podem provocar possíveis heterogeneidades.

Na tabela (5.6) são apresentadas as medidas de posição das variáveis contínuas. Diferentemente do estudo de Rosa et al. (2016), o nível socioeconômico não foi tratado como variável contínua, pois seria um equívoco, uma vez que esta é uma variável categórica. Para uma abordagem contínua dessa informação, é possível utilizar o seu valor absoluto, que também é disponibilizado pelo Inep.

Tabela 5.1 - Distribuição de escola pelo Inse

Inse	N
Muito baixo a baixo	9.116
Médio a muito alto	26.737

Tabela 5.2 - Distribuição de escola pela existência de biblioteca e/ou sala de leitura

Existência de biblioteca e/ou sala de leitura	N
Com biblioteca	26.166
Sem biblioteca	9.687

Tabela 5.3 - Distribuição de escola por localização

Localização	N
Urbano	30.518
Rural	5.335

Tabela 5.4 - Distribuição de escola pelo ICG

ICG	N
Níveis 1 e 2	10.801
Níveis 3 e 4	14.497
Níveis 5 e 6	10.255

Tabela 5.5 - Distribuição de escola por UF

UF	N
RO	419
AC	202
AM	785
RR	85
PA	1.870
AP	177
TO	413
MA	1.483
PI	710
CE	1.429
RN	812

Tabela 5.5 - Distribuição de escola por UF

UF	N
PB	870
PE	1.572
AL	687
SE	505
BA	2.795
MG	3.628
ES	755
RJ	2.128
SP	5.509
PR	2.049
SC	1.511
RS	2.643
MS	575
MT	694
GO	1.219
DF	328

Tabela 5.6 - Medidas de posição das variáveis contínuas

Variável	Mínimo	1° quartil	Mediana	Média	3° quartil	Máximo
Nota média Prova Brasil	139,9	197,7	214,4	213,3	228,6	323,8
Docentes por aluno	0,003	0,042	0,051	0,053	0,061	0,244
Computadores por aluno	0,0004	0,0210	0,0377	0,0471	0,0605	4,5009
Tempo	157,5	240,0	255,0	274,1	270,0	690,0
Histórico escolar	1,30	2,59	2,72	2,70	2,83	3,15
Atenção dos pais	0,25	0,98	1,00	1,00	1,02	1,19
Acesso à cultura	0,04	1,24	1,40	1,41	1,57	3,10
Dedicação extraclasse	0,06	3,13	3,27	3,23	3,38	3,67
AFD	0,60	45,90	65,00	61,44	79,60	100,00
IED	3,10	48,20	63,70	63,66	80,00	100,00

Tabela 5.6 - Medidas de posição das variáveis contínuas

Variável	Mínimo	1° quartil	Mediana	Média	3° quartil	Máximo
IRD	1,0	2,7	3,1	3,1	3,5	4,9

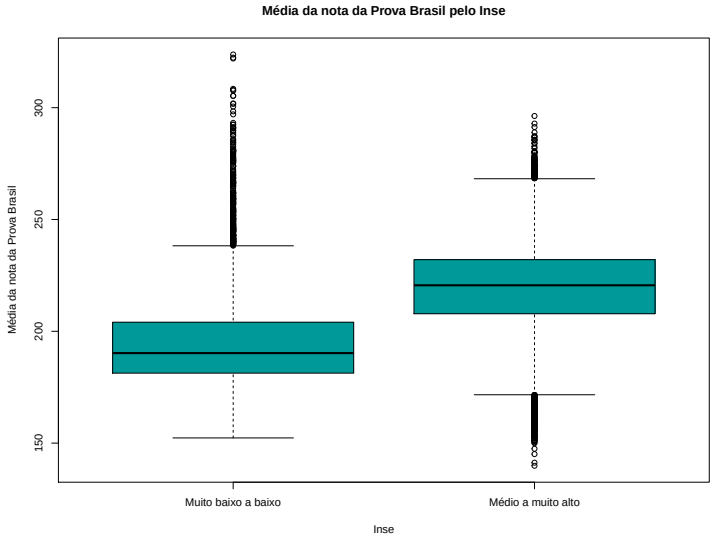

Figura 5.1: Boxplot Inse

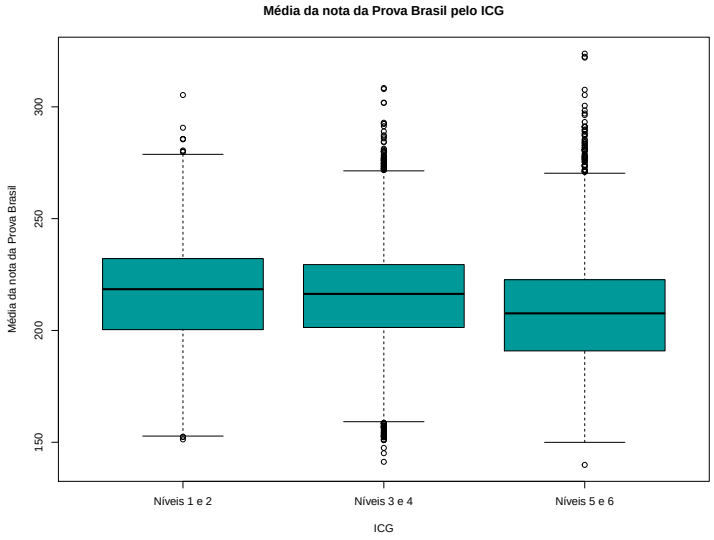

Figura 5.2: Boxplot ICG

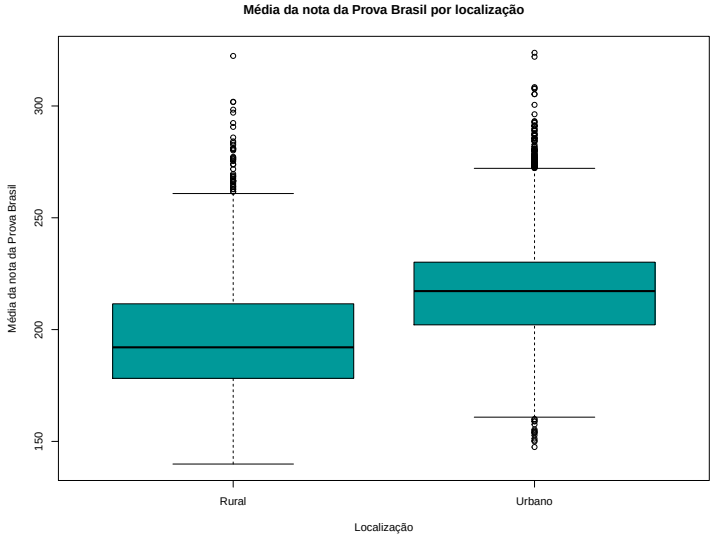

Figura 5.3: Boxplot localização

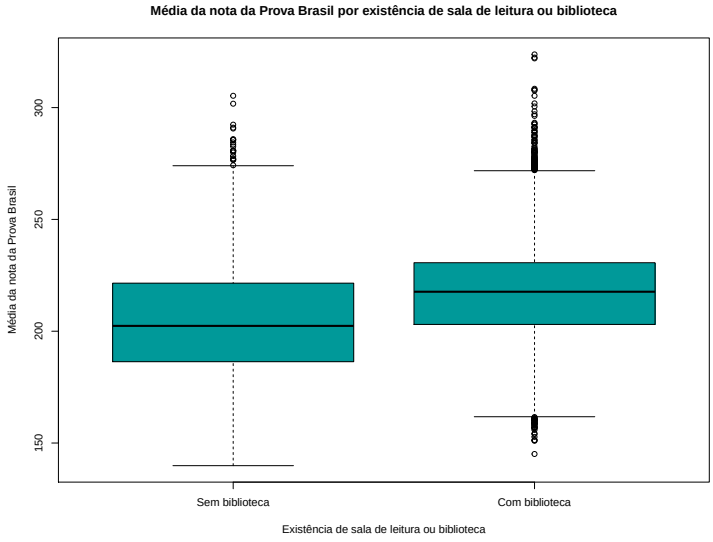

Figura 5.4: Boxplot biblioteca

5.2 Modelo de fronteira estocástica de produção

Witte e López-Torres (2017) apresentaram uma extensa revisão bibliográfica sobre eficiência na educação, incluindo uma revisão das principais variáveis que são comumente usadas para medir eficiência em métodos de fronteira.

Em relação as variáveis relacionadas aos alunos, o desempenho acadêmico prévio e a etnia dos estudantes são variáveis comumente utilizadas.

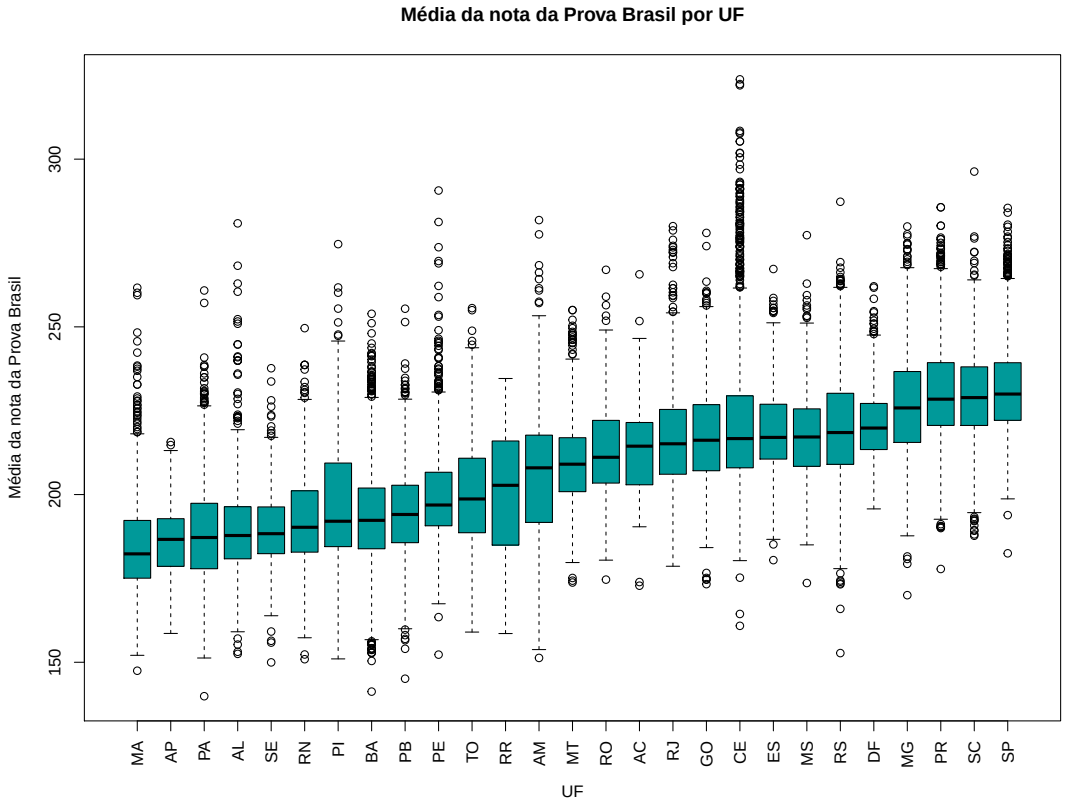

Figura 5.5: Boxplot uf

Em relação ao contexto familiar, o nível socioeconômico das famílias, o nível educacional dos pais e os recursos disponíveis nas residências dos alunos são variáveis frequentemente utilizadas nos estudos.

Já no contexto das escolas, as variáveis mais frequentes são: despesas (com professores, pesquisa, adminsitração), número de professores ou profissionais de outras áreas que atuam na escola, recursos físicos da escola (computadores, livros, salas de aula, construções), número de alunos, número de alunos por professor e experiência do docente em educação.

Em relação ao território em que a escola está instalada, variáveis como localização urbana ou rural, número de escolas no mesmo território e características do local como impostos e taxas de desemprego também são comumente utilizadas nesses estudos.

Por fim, como produto nos modelos de eficiência, resultados em testes de proficiências, número de publicações acadêmicas e número de aprovações podem ser utilizados.

Para definir as variáveis que vão compor o modelo de fronteira estocástica obtido nesta dissertação, além das contribuições de Witte e López-Torres (2017), foram considerados os

modelos desenvolvidos por Rosa et al. (2016) e Trigo (2010). Pois o estudo aqui desenvolvido se assemelha aos dois últimos em relação ao grupo de escolas que está sendo analisado e à prova de proficiência utilizada.

Desta forma, como insumo foram consideradas características físicas da escola. Para modelar a ineficiência, foram consideradas variáveis relacionadas à formação, esforço e experiência do docente, complexidade de gestão escolar, características familiares, características dos alunos e nível socioeconômico da escola, uma vez que acredita-se que a escola não tem o controle desses fatores. Ainda, assim como Rosa et al. (2016), optou-se por controlar a variância do erro bilateral por localização (urbana/rural) e por unidade da federação em que a unidade escolar está instalada.

Desta forma, para a análise da fronteira estocástica de produção, será utilizada uma função de produção do tipo Cobb-Douglas. A ineficiência será modelada por uma distribuição seminormal, isto é, $u \sim \mathrm{N}_{+}(0, \sigma_u^2)$, em que $\sigma_u^2 = \exp(\alpha_u)$, $\alpha_u = m^{'}\delta$, $m^{'}$ é o vetor de variáveis da ineficiência e δ o vetor de parâmetros desconhecidos. Além disso, a variância do erro bilateral será heteroscedástica, ou seja, $v \sim \mathrm{N}(0, \sigma_v^2)$, $\sigma_v^2 = \exp(\alpha_v)$, $\alpha_v = h^{'}w$, $h^{'}$ é o vetor de variáveis com características externas e w o vetor de parâmetros desconhecidos.

O modelo de fronteira estocástica de produção ajustado é dado pela equação (5.1). E os resultados desse modelo são apresentados na tabela (5.7). Para as variáveis de insumo, os resultados apontam que maiores quantidades refletem maiores notas, o que são resultados compatíveis com a literatura. Rosa et al. (2016), Trigo (2010) obtiveram resultados semelhantes.

Em relação à modelagem do termo de ineficiência das escolas, alguns resultados são inesperados. Das variáveis que captam o contexto do aluno e da sua família, o nível socioeconômico (categoria mais elevada), o histórico escolar, a dedicação extraclasse e a atenção dos pais diminuem a ineficiência. Entretanto, o acesso à cultura aumenta a ineficiência. Em relação aos docentes, uma maior proporção de docentes com formação adequada e uma maior regularidade do corpo docente diminuem a ineficiência da escola. Porém, para o indicador de esforço docente o resultado é contraintuitivo, uma vez que o modelo aponta que escolas com menor percentual de docentes nos grupos um e dois do indicador teriam pior desempenho. Por fim, para o indicador de complexidade de gestão da escola, o modelo estimado indica que não há evidências para concluir que os níveis 1 e 2 são distintos, ao contrário do nível 3, e indica que um nível maior de complexidade de

gestão da escola (ICG_3) aumenta a ineficiência.

$$
\begin{aligned}
\mathrm{E}(\ln(\text{Nota média Prova Brasil}_i)|x_i) &= \beta_0 + \beta_1 \ln(\text{Docentes por aluno}_i) \\
&+ \beta_2(\text{Existência de biblioteca\_1}_i) \\
&+ \beta_3 \ln(\text{Computadores por aluno}_i) + \beta_4 \ln(\text{Tempo}_i), \\
\ln(\sigma^2_{ui}) &= \alpha_{u0} + \alpha_{u1}(\text{Inse\_1}_i) + \alpha_{u2} \ln(\text{Histórico escolar}_i) \\
&+ \alpha_{u3} \ln(\text{Atenção dos Pais}_i) + \alpha_{u4} \ln(\text{Acesso à cultura}_i) \\
&+ \alpha_{u5} \ln(\text{Dedicação extraclasse}_i) + \alpha_{u6} \ln(\text{AFD}_i) \\
&+ \alpha_{u7} \ln(\text{IED}_i) + \alpha_{u8}(\text{ICG\_2}_i) + \alpha_{u9}(\text{ICG\_3}_i) + \alpha_{u10} \ln(\text{IRD}_i), \\
\ln(\sigma^2_{vi}) &= \alpha_{v0} + \alpha_{vj} \sum_{j=1}^{26} (\text{UF}_j)_i + \alpha_{v27}(\text{Urbano\_1}_i)
\end{aligned} \tag{5.1}
$$

Tabela 5.7 - Modelo de fronteira estocástica de produção

	Coef	Std.Err	z	P.value	LI	LS
Insumos						
Constante	5,352	0,015	3,682e+02	0,000e+00	5,324	5,381
ln Docentes por aluno	0,006	0,002	3,682e+00	2,314e-04	0,003	0,009
Existência de biblioteca_1	0,025	0,001	2,427e+01	0,000e+00	0,023	0,027
ln Computadores por aluno	0,016	0,001	3,132e+01	0,000e+00	0,015	0,017
ln Tempo	0,023	0,002	9,826e+00	0,000e+00	0,019	0,028
$\ln(\sigma^2_u)$						
Constante	13,886	0,357	3,895e+01	0,000e+00	13,187	14,584
Inse_1	-1,461	0,036	-4,012e+01	0,000e+00	-1,533	-1,390
ln Histórico escolar	-14,566	0,279	-5,219e+01	0,000e+00	-15,113	-14,019
ln Atenção dos pais	-2,506	0,299	-8,395e+00	0,000e+00	-3,092	-1,921
ln Acesso à cultura	0,608	0,086	7,073e+00	1,514e-12	0,440	0,777
ln Dedicação extraclasse	-1,942	0,164	-1,185e+01	0,000e+00	-2,263	-1,621
ln AFD	-0,509	0,025	-2,070e+01	0,000e+00	-0,558	-0,461
ln IED	0,102	0,041	2,524e+00	1,161e-02	0,023	0,182
ICG_2	0,074	0,040	1,831e+00	6,708e-02	-0,005	0,153
ICG_3	0,245	0,043	5,650e+00	1,601e-08	0,160	0,330
ln IRD	-0,481	0,074	-6,519e+00	7,092e-11	-0,625	-0,336

Tabela 5.7 - Modelo de fronteira estocástica de produção

	Coef	Std.Err	z	P.value	LI	LS
$\ln(\sigma_v^2)$						
Constante	-5,909	0,029	-2,046e+02	0,000e+00	-5,965	-5,852
RO	0,398	0,091	4,354e+00	1,335e-05	0,219	0,577
AC	0,108	0,138	7,802e-01	4,353e-01	-0,163	0,378
AM	0,804	0,072	1,124e+01	0,000e+00	0,664	0,944
RR	1,437	0,187	7,686e+00	1,510e-14	1,071	1,804
PA	1,439	0,052	2,793e+01	0,000e+00	1,338	1,540
AP	2,126	0,124	1,715e+01	0,000e+00	1,883	2,369
TO	1,318	0,094	1,409e+01	0,000e+00	1,134	1,501
MA	1,615	0,057	2,822e+01	0,000e+00	1,503	1,727
PI	1,399	0,075	1,862e+01	0,000e+00	1,251	1,546
CE	1,549	0,045	3,470e+01	0,000e+00	1,462	1,637
RN	1,901	0,060	3,176e+01	0,000e+00	1,783	2,018
PB	1,578	0,065	2,430e+01	0,000e+00	1,451	1,705
PE	0,944	0,056	1,688e+01	0,000e+00	0,834	1,053
AL	1,638	0,072	2,269e+01	0,000e+00	1,496	1,779
SE	1,618	0,081	1,990e+01	0,000e+00	1,459	1,778
BA	1,490	0,040	3,727e+01	0,000e+00	1,412	1,568
MG	0,812	0,031	2,624e+01	0,000e+00	0,751	0,872
ES	0,379	0,064	5,908e+00	3,461e-09	0,253	0,504
RJ	0,355	0,043	8,291e+00	2,220e-16	0,271	0,439
SP	0,705	0,029	2,452e+01	0,000e+00	0,649	0,761
PR	0,915	0,038	2,425e+01	0,000e+00	0,842	0,989
SC	0,717	0,042	1,696e+01	0,000e+00	0,634	0,800
RS	0,674	0,036	1,890e+01	0,000e+00	0,605	0,744
MS	0,461	0,073	6,324e+00	2,551e-10	0,318	0,605
MT	1,314	0,060	2,196e+01	0,000e+00	1,196	1,431
GO	0,620	0,051	1,209e+01	0,000e+00	0,520	0,721
Urbano_1	-0,507	0,030	-1,694e+01	0,000e+00	-0,566	-0,449

5.3 Estimação em dois estágios e correção de Murphy Topel

Uma aplicação comum do método de dois estágios é contabilizar, no segundo estágio, a variação de um regressor construído. Nesse sentido, a variável construída tipicamente é uma estimativa do valor esperado de uma variável que provavelmente será endógena no modelo do segundo estágio (Greene (1990)). Neste aspecto, a estimação em dois estágios empresta ao processo propriedades robustas relativamente a endogeneidade.

Além disso, em caso de bases de dados com informações ausentes, por exemplo, o uso do valor predito de um indicador predito (probabilidade de sucesso) no modelo de fronteira estocástica potencialmente aumenta o tamanho da amostra.

Considerando a natureza das variáveis utilizadas no modelo de fronteira estocástica de produção nesta dissertação, possivelmente o indicador do nível socioeconômico é uma variável endógena. Entretanto, não foi objeto desta dissertação aprofundar o estudo da endogeneidade. O exercício aqui aplicado se limitou ao processo em dois estágios e à correção de Murphy-Topel.

O primeiro estágio é ajustado por um modelo de regressão logística, cuja variável resposta é o Inse. O modelo estimado é dado pela equação (5.2) e seus resultados são apresentados na tabela (5.8). Neste ponto, faz-se uma observação. Esta análise seria executada no software *Stata*, porém o modelo logístico não convergiu quando inseridas as variáveis dicotômicas que representam as unidades da federação. Caso a análise fosse feita por região, o software poderia ter sido utilizado. Desta forma, optou-se pelo R.

$$
\begin{aligned}
\mathrm{P}(\mathrm{Inse}_i = 1|x_i) &= \mathrm{logit}^{-1}[\theta_0 + \theta_1 \ln(\text{Docentes por aluno}_i) + \theta_2(\text{Existência de biblioteca}_i) \\
&+ \theta_3 \ln(\text{Computadores por aluno}_i) + \theta_4 \ln(\text{Tempo}_i) + \theta_5(\text{Urbano\_1}_i) \\
&+ \theta_6 \ln(\text{Histórico escolar}_i) + \theta_7 \ln(\text{Atenção dos Pais}_i) + \theta_8 \ln(\text{Acesso à cultura}_i) \\
&+ \theta_9 \ln(\text{Dedicação extraclasse}_i) + \theta_{10} \ln(\mathrm{AFD}_i) + \theta_{11} \ln(\mathrm{IED}_i) \\
&+ \theta_{12}(\mathrm{ICG\_2}_i) + \theta_{13}(\mathrm{ICG\_3}_i) + \theta_{14} \ln(\mathrm{IRD}_i) + \theta_j \sum_{j=15}^{40} (\mathrm{UF}_{ji})] \qquad (5.2)
\end{aligned}
$$

Tabela 5.8 - Primeiro estágio: modelo de regressão logístico

	Coef	Std.Err	z	P.value	LI	LS

Tabela 5.8 - Primeiro estágio: modelo de regressão logístico

Constante	9,046	81,741	1,107e-01	9,119e-01	-151,163	169,254
ln Docentes por aluno	-0,331	0,066	-5,031e+00	4,869e-07	-0,459	-0,202
Existência de biblioteca	0,298	0,039	7,572e+00	3,686e-14	0,221	0,375
ln Computadores por aluno	-0,047	0,018	-2,559e+00	1,051e-02	-0,083	-0,011
ln Tempo	0,410	0,097	4,238e+00	2,251e-05	0,221	0,600
Urbano_1	1,213	0,046	2,659e+01	0,000e+00	1,124	1,303
ln Histórico Escolar	-1,750	0,318	-5,503e+00	3,724e-08	-2,373	-1,127
ln Atenção dos pais	2,428	0,335	7,242e+00	4,405e-13	1,771	3,085
ln Acesso à cultura	-1,429	0,098	-1,458e+01	0,000e+00	-1,621	-1,237
ln Dedicação extraclasse	0,834	0,194	4,289e+00	1,799e-05	0,453	1,214
ln AFD	0,579	0,032	1,790e+01	0,000e+00	0,516	0,642
ln IED	-0,088	0,047	-1,887e+00	5,921e-02	-0,179	0,003
ICG_2	-0,311	0,046	-6,814e+00	9,512e-12	-0,401	-0,222
ICG_3	-0,412	0,048	-8,516e+00	0,000e+00	-0,507	-0,317
ln IRD	0,137	0,085	1,623e+00	1,046e-01	-0,028	0,303
RO	-12,838	81,737	-1,571e-01	8,752e-01	-173,040	147,364
AC	-13,670	81,737	-1,672e-01	8,672e-01	-173,872	146,532
AM	-13,407	81,737	-1,640e-01	8,697e-01	-173,608	146,795
RR	-12,152	81,738	-1,487e-01	8,818e-01	-172,355	148,051
PA	-14,467	81,737	-1,770e-01	8,595e-01	-174,669	145,734
AP	-12,700	81,737	-1,554e-01	8,765e-01	-172,903	147,502
TO	-14,139	81,737	-1,730e-01	8,627e-01	-174,341	146,063
MA	-14,916	81,737	-1,825e-01	8,552e-01	-175,118	145,286
PI	-15,013	81,737	-1,837e-01	8,543e-01	-175,215	145,189
CE	-15,025	81,737	-1,838e-01	8,542e-01	-175,227	145,176
RN	-14,275	81,737	-1,746e-01	8,614e-01	-174,477	145,927
PB	-14,855	81,737	-1,817e-01	8,558e-01	-175,057	145,347
PE	-14,200	81,737	-1,737e-01	8,621e-01	-174,401	146,002
AL	-14,923	81,737	-1,826e-01	8,551e-01	-175,125	145,279
SE	-14,862	81,737	-1,818e-01	8,557e-01	-175,064	145,340

Tabela 5.8 - Primeiro estágio: modelo de regressão logístico

BA	-14,340	81,737	-1,754e-01	8,607e-01	-174,542	145,862
MG	-12,314	81,737	-1,506e-01	8,803e-01	-172,515	147,888
ES	-11,865	81,737	-1,452e-01	8,846e-01	-172,067	148,337
RJ	-8,079	81,739	-9,884e-02	9,213e-01	-168,284	152,126
SP	-9,595	81,737	-1,174e-01	9,066e-01	-169,797	150,607
PR	-10,566	81,737	-1,293e-01	8,971e-01	-170,768	149,636
SC	-8,077	81,739	-9,881e-02	9,213e-01	-168,282	152,129
RS	-7,438	81,739	-9,100e-02	9,275e-01	-167,644	152,768
MS	-10,659	81,738	-1,304e-01	8,962e-01	-170,862	149,544
MT	-11,684	81,737	-1,430e-01	8,863e-01	-171,886	148,518
GO	-10,879	81,737	-1,331e-01	8,941e-01	-171,081	149,323

O segundo estágio é a fronteira estocástica de produção de interesse. A equação do modelo do segundo estágio ajustado é a mesma equação em (5.1), com uma distinção. A variável Inse é agora substituída pela probabilidade predita pelo modelo logístico. A nova variável é denominada P_Inse e indica a probabilidade da escola pertencer aos grupos médio a muito alto do indicador.

Como explicado no capítulo 3, as estimativas de erro padrão calculadas na estimação em dois estágios são inconsistentes. Por isso, foi calculada a correção de Murphy-Topel, apresentada junto aos resultados do segundo estágio na tabela 5.9. Para comparação entre as estimativas de erro padrão com e sem a correção, a tabela 5.10 apresenta-os com um número maior de casas decimais. Em relação aos resultados, o modelo de fronteira estocástica de produção ajustado no segundo estágio apresentou resultados semelhantes ao modelo ajustado em (5.7). Em relação aos insumos, verifica-se uma relação positiva entre o desempenho da escola na Prova Brasil e o aumento dos insumos.

Na modelagem do termo de ineficiência das escolas, destaca-se que a probabilidade da escola pertencer a um contexto onde o nível socioeconômico é mais alto e níveis maiores do histórico escolar do aluno, da dedicação extraclasse do aluno e da atenção dos pais diminuem a ineficiência. O acesso à cultura não é uma variável significativa no modelo.

Em relação aos docentes, uma maior proporção de docentes com formação adequada e uma maior regularidade do corpo docente diminuem a ineficiência da escola. Porém,

para o indicador de esforço docente o resultado é contraintuitivo, uma vez que o modelo aponta que escolas com menor percentual de docentes nos grupos um e dois do indicador (docentes com menor sobrecarga) teriam pior desempenho. Rosa et al. (2016) também encontra resultado semelhante e aponta a importância de interpretar esses resultados com cautela, uma vez que a sobrecarga de professores pode ser prejudicial aos mesmos.

Por fim, o indicador de complexidade de gestão da escola também tem um resultado inesperado, o modelo estimado indica que não há evidências para concluir que os níveis 1 e 3 são distintos, ao contrário do nível 2, e indica que um nível maior de complexidade de gestão da escola (ICG_3) diminui menos a ineficiência do que um nível menor.

Tabela 5.9 - Segundo estágio: modelo de fronteira estocástica de produção com correção de Murphy-Topel

	Coef	Std.Err	z	P.value	LI	LS
Insumos						
Constante	5,362	0,016	3,366e+02	0,000e+00	5,331	5,393
ln Docentes por alunos	0,006	0,002	3,749e+00	1,776e-04	0,003	0,009
Existência de biblioteca	0,018	0,001	1,650e+01	0,000e+00	0,016	0,020
ln Computadores por aluno	0,016	0,001	3,047e+01	0,000e+00	0,015	0,017
ln Tempo	0,023	0,003	8,700e+00	0,000e+00	0,018	0,028
$\ln(\sigma_u^2)$						
Constante	10,600	0,501	2,115e+01	0,000e+00	9,618	11,582
P_Inse	-3,241	0,169	-1,921e+01	0,000e+00	-3,571	-2,910
ln Histórico escolar	-11,425	0,452	-2,527e+01	0,000e+00	-12,312	-10,539
ln Atenção dos pais	-0,773	0,263	-2,941e+00	3,274e-03	-1,289	-0,258
ln Acesso à cultura	-0,169	0,112	-1,512e+00	1,305e-01	-0,389	0,050
ln Dedicação extraclasse	-1,829	0,127	-1,442e+01	0,000e+00	-2,078	-1,580
ln AFD	-0,081	0,031	-2,597e+00	9,417e-03	-0,143	-0,020
ln IED	0,151	0,037	4,079e+00	4,526e-05	0,078	0,223
ICG_2	-0,127	0,034	-3,723e+00	1,965e-04	-0,193	-0,060
ICG_3	-0,033	0,043	-7,619e-01	4,461e-01	-0,118	0,052
ln IRD	-0,294	0,063	-4,642e+00	3,443e-06	-0,418	-0,170
$\ln(\sigma_v^2)$						
Constante	-6,162	0,046	-1,337e+02	0,000e+00	-6,252	-6,072

Tabela 5.9 - Segundo estágio: modelo de fronteira estocástica de produção com correção de Murphy-Topel

	Coef	Std.Err	z	P.value	LI	LS
RO	0,247	0,102	2,410e+00	1,595e-02	0,046	0,448
AC	0,031	0,154	2,009e-01	8,408e-01	-0,271	0,333
AM	0,805	0,078	1,034e+01	0,000e+00	0,653	0,958
RR	1,552	0,188	8,252e+00	2,220e-16	1,183	1,920
PA	0,627	0,106	5,922e+00	3,180e-09	0,419	0,834
AP	-18,360	0,000	-2,079e+09	0,000e+00	-18,360	-18,360
TO	1,020	0,117	8,715e+00	0,000e+00	0,790	1,249
MA	0,810	0,113	7,147e+00	8,853e-13	0,588	1,033
PI	0,888	0,107	8,309e+00	0,000e+00	0,679	1,098
CE	1,730	0,054	3,201e+01	0,000e+00	1,624	1,836
RN	1,554	0,073	2,123e+01	0,000e+00	1,411	1,698
PB	0,618	0,136	4,556e+00	5,215e-06	0,352	0,884
PE	0,534	0,073	7,290e+00	3,104e-13	0,391	0,678
AL	1,074	0,114	9,419e+00	0,000e+00	0,851	1,298
SE	0,769	0,172	4,475e+00	7,633e-06	0,432	1,106
BA	0,840	0,069	1,212e+01	0,000e+00	0,704	0,975
MG	0,845	0,035	2,422e+01	0,000e+00	0,776	0,913
ES	0,313	0,069	4,563e+00	5,054e-06	0,179	0,448
RJ	0,392	0,045	8,633e+00	0,000e+00	0,303	0,481
SP	0,673	0,032	2,087e+01	0,000e+00	0,610	0,736
PR	0,893	0,042	2,111e+01	0,000e+00	0,810	0,976
SC	0,701	0,045	1,562e+01	0,000e+00	0,613	0,789
RS	0,722	0,038	1,878e+01	0,000e+00	0,647	0,798
MS	0,462	0,076	6,091e+00	1,121e-09	0,314	0,611
MT	1,318	0,062	2,142e+01	0,000e+00	1,197	1,438
GO	0,631	0,054	1,168e+01	0,000e+00	0,525	0,736
Urbano_1	-0,217	0,039	-5,569e+00	2,566e-08	-0,294	-0,141

Tabela 5.10 - Estimativas de erro-padrão com e sem a correção de Murphy-Topel

	Coef	Std.Err com MT	Std.Err sem MT
Insumos			
Constante	5,36220	0,01593	0,01462
ln Docentes por alunos	0,00616	0,00164	0,00153
Existência de biblioteca	0,01789	0,00108	0,00104
ln Computadores por aluno	0,01611	0,00053	0,00052
ln Tempo	0,02296	0,00264	0,00237
$\ln(\sigma_u^2)$			
Constante	10,60003	0,50114	0,29798
P_Inse	-3,24057	0,16871	0,05854
ln Histórico escolar	-11,42534	0,45220	0,22661
ln Atenção dos pais	-0,77326	0,26294	0,23403
ln Acesso à cultura	-0,16928	0,11194	0,06598
ln Dedicação extraclasse	-1,82891	0,12685	0,14971
ln AFD	-0,08122	0,03128	0,02215
ln IED	0,15072	0,03695	0,03215
ICG_2	-0,12666	0,03402	0,03169
ICG_3	-0,03299	0,04330	0,03393
ln IRD	-0,29424	0,06338	0,05817
$\ln(\sigma_v^2)$			
Constante	-6,16189	0,04608	0,03928
RO	0,24695	0,10247	0,10138
AC	0,03092	0,15394	0,15389
AM	0,80544	0,07787	0,07933
RR	1,55177	0,18804	0,18837
PA	0,62657	0,10580	0,10474
AP	-18,36000	0,00000	0,00000
TO	1,01968	0,11700	0,11985
MA	0,81044	0,11339	0,11384
PI	0,88846	0,10692	0,10627

Tabela 5.10 - Estimativas de erro-padrão com e sem a correção de Murphy-Topel

	Coef	Std.Err com MT	Std.Err sem MT
CE	1,72995	0,05405	0,05123
RN	1,55419	0,07321	0,07522
PB	0,61830	0,13571	0,13051
PE	0,53411	0,07327	0,07317
AL	1,07443	0,11406	0,11143
SE	0,76920	0,17188	0,16685
BA	0,83967	0,06928	0,06584
MG	0,84461	0,03487	0,03367
ES	0,31343	0,06870	0,06768
RJ	0,39172	0,04538	0,04512
SP	0,67266	0,03222	0,03165
PR	0,89343	0,04232	0,04000
SC	0,70067	0,04485	0,04430
RS	0,72233	0,03846	0,03818
MS	0,46234	0,07590	0,07581
MT	1,31765	0,06151	0,06191
GO	0,63062	0,05398	0,05384
Urbano_1	-0,21734	0,03903	0,03884

5.4 Eficiência técnica

A eficiência técnica foi calculada conforme a equação (2.19). Considerando o conjunto de escolas analisados, a eficiência técnica varia de 0,65 a 0,99. Estes valores devem ser interpretados dentro desse conjunto de escolas, no contexto das variáveis que foram apresentadas. A tabela (5.11) e as figuras (5.6), (5.7) indicam que a distribuição de probabilidade da eficiência técnica é assimétrica à esquerda.

Tabela 5.11 - Medidas de posição eficiência técnica

Medidas de posição	
Mínimo	0,6515

Tabela 5.11 - Medidas de posição eficiência técnica

Medidas de posição	
1° quartil	0,9012
Mediana	0,9583
Média	0,9319
3° quartil	0,9759
Máximo	0,9910

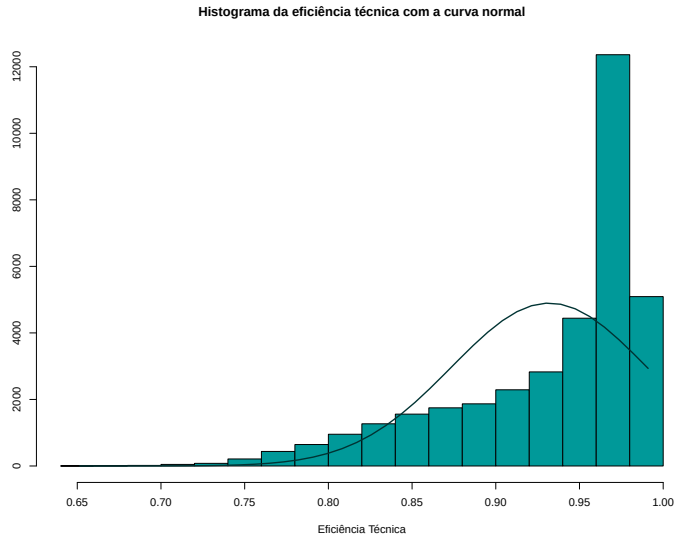

Figura 5.6: Histograma da eficiência técnica

A figura 5.8 indica que a relação entre a eficiência técnica calculada e a nota média obtida na Prova Brasil é positiva, porém não é linear.

Por fim, considerando as variáveis que compõe a variância do erro bilateral, observa-se que a eficiência técnica de escolas na área urbana é maior do que de escolas rurais. Em relação aos resultados da eficiência técnica por UF (5.10), observa-se que as medianas das notas da Prova Brasil do estados do Norte e do Nordeste são mais baixas do que as das demais regiões. Além disso, em geral, a distância entre as notas do 3º quartil e do 1º quartil do Norte e Nordeste também são maiores do que das regiões Sudeste, Sul e Centro-Oeste, ou seja, a eficiência técnica das escolas do Norte e Nordeste é mais heterogênea do que das demais regiões brasileiras.

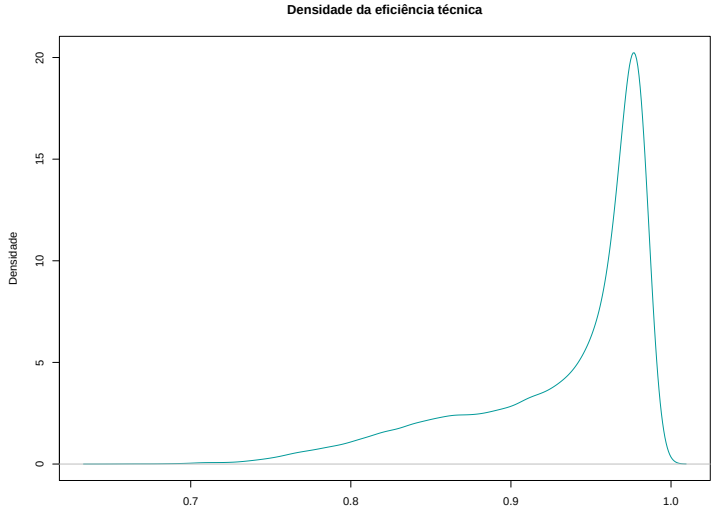

Figura 5.7: Densidade da eficiência técnica

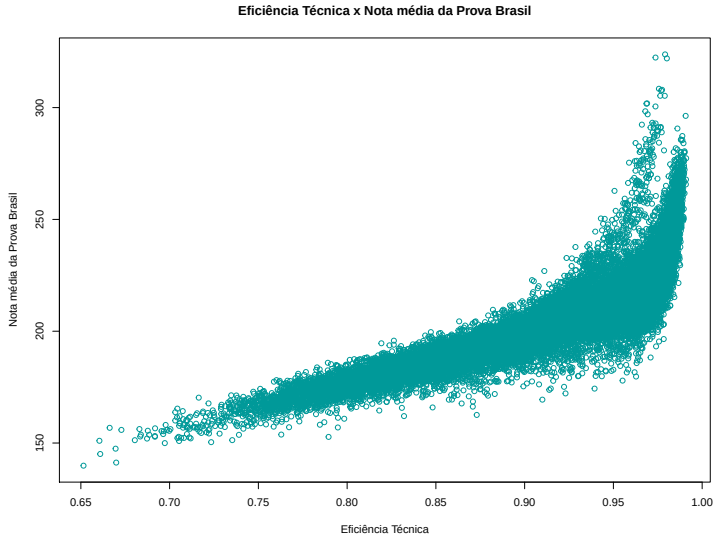

Figura 5.8: Eficiência técnica x Nota média da Prova Brasil

Os resultados aqui encontrados para a eficiência técnica são semelhantes aos apresentados por Rosa et al. (2016).

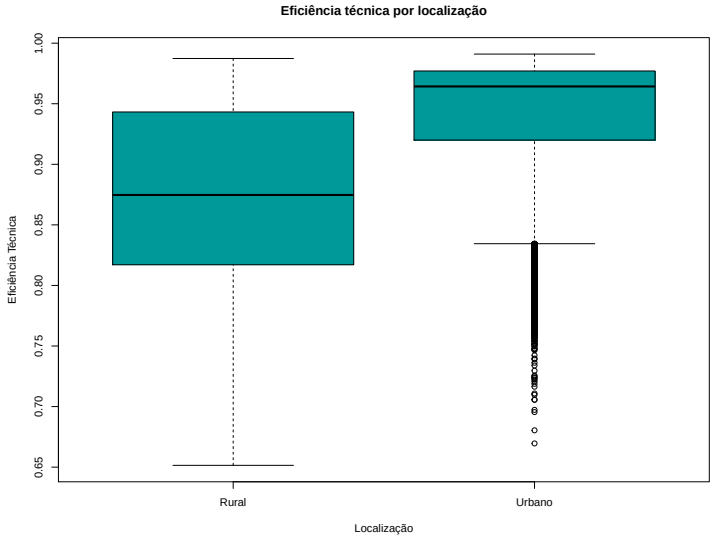

Figura 5.9: Eficiência técnica por localização

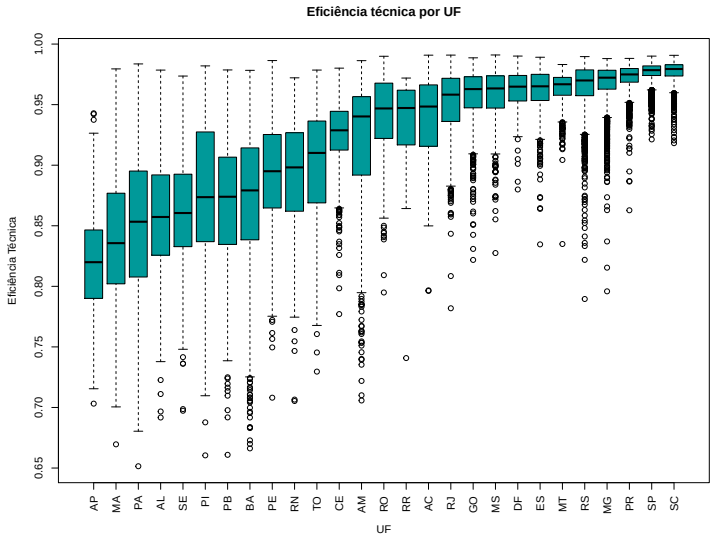

Figura 5.10: Eficiência técnica por UF

Capítulo 6

Conclusão

Este trabalho produziu uma estimação em dois estágios da eficiência técnica das escolas brasileiras a partir dos dados de Censo Escolar 2015 e da Prova Brasil do mesmo ano. Foi considerado como produto das escolas a proficiência dos alunos, a partir das médias das notas de português e matemática. Como insumos, foram considerados recursos físicos. E para ineficiência foram consideradas variáveis que captassem o contexto socioeconômico em que os alunos da escola estão inseridos, além de aspectos da gestão escolar e dos docentes que nela atuam. Considerou-se, ainda, a heterogeneidade que fatores como localização e UF podem ocasionar.

O trabalho de Rosa et al. (2016) foi tomado como base e estendido. Foram abordadas a estimação em dois estágios e a correção de Murphy-Topel para matriz de variância-covariância do modelo do segundo estágio. Em geral, os novos indicadores do Inep testados apresentaram comportamento semelhante àqueles construídos por Rosa et al. (2016), sendo possível reduzir o número de variáveis do modelo.

O modelo estimado no segundo estágio indica que uma quantidade maior de recursos físicos está associada a melhores resultados. Em relação à ineficiência, o histórico escolar do aluno é relevante, assim como o nível socioeconômico em que ele está inserido, a sua dedicação extraclasse e a atenção que os pais dedicam a ele. Como afirmaram Rosa et al. (2016), estes resultados apontam a importância do planejamento de políticas públicas voltadas para aproximação entre família, escola e comunidade. Em relação aos professores, a regularidade do corpo docente na escola se destacou mais do que a formação ou a sobrecarga de trabalho do mesmo. Já a variável de complexidade de gestão da escola apresentou um resultado inesperado, que deverá ser melhor explorado em outros estudos.

Adicionalmente aos resultados educacionais, esta dissertação traz como contribuição o

desenvolvimento de um algoritmo no software R, disponível no apêndice G, que possibilita estudar modelos de fronteira estocástica de produção. O algoritmo considera a questão da estimação em dois estágios e a correção de Murphy-Topel para matriz de variância-covariância do modelo do segundo estágio. Como trabalhos futuros, sugere-se replicar este estudo para o 9º ano do Ensino Fundamental, produzir uma análise longitudinal da eficiência técnica das escolas e desenvolver um pacote no R que contemple o algoritmo aqui desenvolvido.

Referências Bibliográficas

Aigner D., Lovell C., Schmidt P., Formulation and estimation of stochastic frontier production function models, Journal of Econometrics, 1977, vol. 6, p. 21

Andrade B. B., Souza G. S., Likelihood computation in the normal-gamma stochastic frontier model, Computational Statistics, 2018, vol. 33, p. 967

Beckers D. E., Hammond C. J., A tractable likelihood function for the normal-gamma stochastic frontier model, Economics Letters, 1987, vol. 24, p. 33

Bogetoft P., Otto L., Benchmarking with DEA, SFA, and R. International Series in Operations Research & Management Science, Springer New York, 2010

Coelli T., Rao D., O'Donnell C., Battese G., An Introduction to Efficiency and Productivity Analysis. Springer US, 2005

Greene W., Econometric Analysis. Pearson series in economics, Prentice Hall, 2012

Greene W. H., A Gamma-distributed stochastic frontier model, Journal of Econometrics, 1990, vol. 46, p. 141

Hardin J. W., The robust variance estimator for two-stage models, Stata Journal, 2002, vol. 2, p. 253

Holgado-Tello F., Moscoso S., Barbero–García I., Vila E., Polychoric versus Pearson correlations in Exploratory and Confirmatory Factor Analysis with ordinal variables, Quality and Quantity, 2010, vol. 44, p. 153

Inep, 2014a Technical report Indicador de adequação da formação do docente da educação básica. Brasilia, DF

Inep, 2014b Technical report Indicador de Esforço Docente. Brasilia, DF

Inep, 2014c Technical report Indicador de nível socioeconômico (Inse) das escolas. Brasilia, DF

Inep, 2014d Technical report Indicador para mensurar a complexidade da gestão nas escolas a partir dos dados do Censo Escolar da Educação Básica. Brasilia, DF

Inep, 2015 Technical report Indicador de regularidade do docente da Educação Básica. Brasilia, DF

Inep, 2017 Technical report Microdados da Aneb e da Anresc 2015. [online]. Brasília: Inep, 2017. [citado 2017-01-30]. Disponível em: http://portal.inep.gov.br/basica-levantamentos-acessar

Kumbhakar S., Lovell C., Stochastic Frontier Analysis. Stochastic Frontier Analysis, Cambridge University Press, 2003

Meeusen W., van Den Broeck J., Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error, International Economic Review, 1977, vol. 18, p. 435

Murphy K., Topel R., Estimation and Inference in Two-Step Econometric Models, Journal of Business Economic Statistics, 2002, vol. 20, p. 88

Rosa T. M., de Oliveira Gonçalves F., de Andrade K. R., de Moraes Santos T. V., Eficiência técnica nas escolas públicas brasileiras: a situação do Distrito Federal no contexto nacional, 2016

Stevenson R. E., Likelihood functions for generalized stochastic frontier estimation, Journal of Econometrics, 1980, vol. 13, p. 57

Trigo P. P., Avaliação da eficiência técnica do ensino básico brasileiro, 2010

Witte K. D., López-Torres L., Efficiency in education: a review of literature and a way forward, Journal of the Operational Research Society, 2017, vol. 68, p. 339

Apêndice

Apêndice A

Índices contextuais

As variáveis categóricas listadas na tabela A.1 foram utilizadas para construção dos índices: histórico escolar do aluno, acesso à cultura, dedicação extraclasse e atenção dos pais. As categorias de respostas dessas variáveis são descritas no dicionário de variáveis disponibilizado junto aos microdados da Prova Brasil (Inep (2017)).

Tabela A.1 - Variáveis do questionário do aluno do Saeb utilizadas para o cálculos dos índices contextuais

Índice	Composição dos índices
Índice de histórico escolar do aluno	q043 - Quando você entrou na escola? q044 - A partir da primeira série ou primeiro ano, em que tipo de escola você estudou? q045 - Você já foi reprovado? q046 - Você já abandonou a escola durante o período de aulas e ficou fora da escola o resto do ano?
Índice de acesso à cultura	Com qual frequência você lê: q032 - Jornais q033 - Livros q034 - Revistas em geral q035 - Revistas em quadrinho q036 - Notícias na internet Com qual frequência você costuma ir à/ao: q037 - Biblioteca q038 - Cinema q039 - Espetáculo ou exposição

Tabela A.1 - Variáveis do questionário do aluno do Saeb utilizadas para o cálculos dos índices contextuais

Índice	Composição dos índices
Índice dedicação extraclasse	q040 - Em dia de aula, quanto tempo você gasta assistindo à TV, navegando na internet ou jogando jogos eletrônicos? q041 - Em dias de aula, quanto tempo você gasta fazendo trabalhos domésticos? q042 - Atualmente você trabalha fora de casa? q047 - Você faz o dever de casa de Língua Portuguesa? q048 - O(A) professor(a) corrige o dever de casa de Língua Portuguesa? q049 - Você faz o dever de casa de Matemática? q050 - O(A) professor(a) corrige o dever de casa de Matemática? q051 - Você utiliza a biblioteca ou sala de leitura da sua escola?
Índice de atenção dos pais	q021 - Você vê sua mãe, ou mulher responsável por você, lendo? q025 - Você vê o seu pai, ou homem responsável por você, lendo? q027 - Seus pais ou responsáveis incentivam você a estudar? q028 - Seus pais ou responsáveis incentivam você a fazer o dever de casa e/ou os trabalhos da escola? q029 - Seus pais ou responsáveis incentivam você a ler? q030 - Seus pais ou responsáveis incentivam você a ir a escola e/ou não faltar às aulas? q031 - Seus pais ou responsáveis conversam com você sobre o que acontece na escola?

Apêndice B

Condições de regularidade

Definição 1. Condições de regularidade

1. As primeiras três derivadas de $\ln f(y_i|\beta)$ com relação a β são contínuas e finitas para quase todos y_i e para todo β. Esta condição garante a existência de uma certa aproximação de série de taylor e a variância finita das derivadas de $\ln \mathrm{L}$.

2. As condições necessárias para obter as esperanças da primeira e segunda derivadas de $\ln f(y_i|\beta)$ estão satisfeitas.

3. Para todos os valores de β, $|\partial^3 \ln f(y_i|\beta)/\partial\beta_j\partial\beta_k\partial\beta_l|$ é menor que uma função que tem esperança finita. Essa condições permite truncar a série de Taylor.

Apêndice C

Momentos de derivadas de log-verossimilhança

Densidades regulares apresentam três propriedades.

Teorema 1. Momentos de derivadas de log-verossimilhança

1. $\ln f(y_i|\beta)$, $g_i = \partial \ln f(y_i|\beta)/\partial\beta$, e $\mathrm{H}_i = \partial^2 \ln f(y_i|\beta)/\partial\beta\partial\beta'$, $i = 1, ..., n$ são amostras aleatórias de variáveis aleatórias. A notação $g_i(\beta_0)$ e $\mathrm{H}_i(\beta_0)$ indica a derivada no ponto β_0.

2. $\mathrm{E}_0[g_i(\beta_0)] = 0$

3. $\mathrm{Var}[g_i(\beta_0)] = -\mathrm{E}[\mathrm{H}_i(\beta_0)]$

Apêndice D

Igualdade da matriz de informação

A Hessiana da log-verossimilhança é

$$\mathrm{H} = \frac{\partial^2 \ln L(\beta|y)}{\partial\beta\partial\beta'} = \sum_{i=1}^{n} \frac{\partial^2 \ln f(y_i|\beta)}{\partial\beta\partial\beta'} = \sum_{i=1}^{n} \mathrm{H}_i$$

Evoluindo em β_0, temos

$$\mathrm{E}_0[g_0 g_0^{'}] = \mathrm{E}_0 \left[\sum_{i=1}^{n} \sum_{j=1}^{n} g_{0i} g_{0j}^{'} \right],$$

e, por causa de D_1[1], eliminando termos com subscritos diferentes, obtém-se

$$\mathrm{E}_0[g_0 g_0^{'}] = \mathrm{E}_0 \left[\sum_{i=1}^{n} \sum_{j=1}^{n} g_{0i} g_{0i}^{'} \right] = \mathrm{E}_0 \left[\sum_{i=1}^{n} (-\mathrm{H}_{0i}) \right] = -\mathrm{E}_0 \left[\mathrm{H}_0 \right],$$

logo

$$\mathrm{Var}_0 \left[\frac{\partial \ln L(\beta_0|y)}{\partial\beta_0} \right] = \mathrm{E}_0 \left[\left(\frac{\partial \ln L(\beta_0|y)}{\partial\beta_0} \right) \left(\frac{\partial \ln L(\beta_0|y)}{\partial\beta_0'} \right) \right] = -\mathrm{E}_0 \left[\frac{\partial^2 \ln L(\beta_0|y)}{\partial\beta_0\partial\beta_0'} \right]. \quad \text{(D.1)}$$

Esse resultado é conhecido como matriz de informação de igualdade.

[1] $\ln f(y_i|\beta)$, $g_i = \partial \ln f(y_i|\beta)/\partial\beta$, e $\mathrm{H}_i = \partial^2 \ln f(y_i|\beta)/\partial\beta\partial\beta^{'}$, $i = 1, ..., n$ são amostras aleatórias de variáveis aleatórias. A notação $g_i(\beta_0)$ e $\mathrm{H}_i(\beta_0)$ indica a derivada no ponto β_0.

Apêndice E

Normalidade assintótica

No estimador de máxima verossimilhança, o gradiente da log-verossimilhança é igual a zero por definição, então

$$g(\hat{\beta}) = 0.$$

Expandindo essas equações em uma série de Taylor, obtém-se:

$$g(\hat{\beta}) = g(\beta_0) + \mathrm{H}(\bar{\beta})(\hat{\beta} - \beta_0) = 0.$$

Evolui-se a hessiana até o ponto $\bar{\beta}$ que está entre $\hat{\beta}$ e $\beta_0[\bar{\beta} = w\hat{\beta} + (1-w)\beta_0$ para $0 < w < 1]$. Rearranjando esta função e multiplicando por $\sqrt{n}$, tem-se:

$$\sqrt{n}(\hat{\beta} - \beta_0) = [-\mathrm{H}(\bar{\beta})]^{-1}[\sqrt{n}g(\beta_0)].$$

Como $(\hat{\beta} - \beta_0) = 0$, então $(\hat{\beta} - \bar{\beta_0})$ As derivadas segundas são funções contínuas. Portanto, se a distribuição limite existe, então

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} [-\mathrm{H}(\beta)]^{-1}[\sqrt{n}g(\beta_0)].$$

Dividindo $\mathrm{H}(\beta_0)$ e $g(\beta_0)$ por n, obtém-se

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} [-1/n\mathrm{H}(\beta)]^{-1}[\sqrt{n}\bar{g}(\beta_0)]. \tag{E.1}$$

Aplicando o teorema do limite central de Lindeberg-Levy, a variância limite de $[\sqrt{n}\bar{g}(\beta_0)]$ é $-\mathrm{E}[1/n\mathrm{H}(\beta_0)]$, então

$$\sqrt{n}\bar{g}(\beta_0) \xrightarrow{d} N[0, -\mathrm{E}_0[1/n\mathrm{H}(\beta_0)]].$$

Por causa da segunda propriedade do teorema do apêndice C, $plim[-1/n\mathrm{H}(\beta_0)] = -\mathrm{E}_0[1/n\mathrm{H}(\beta_0)]$. Isto resulta em uma matriz constante, que permite obter o resultado

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N[0, \{-\mathrm{E}_0[1/n\mathrm{H}(\beta_0)]\}^{-1}],$$

que fornece a distribuição assintótica do estimador de máxima verossimilhança:

$$\hat{\beta} \sim N\left[\beta_0, \{\mathrm{I}(\beta_0)\}^{-1}\right].$$

Apêndice F

Correlação policórica

Suponha que Z_1 e Z_2 são duas variáveis ordinais com m_1 e m_2 categorias. A sua distribuição na amostra é dada conforme a tabela de contingência

$$\begin{array}{cccc} n_{11} & n_{12} & ... & n_{1m2} \\ n_{21} & n_{22} & ... & n_{2m2} \\ . & . & . & . \\ . & . & n_{ij} & . \\ . & . & . & . \\ n_{m11} & n_{m11} & ... & n_{m1m2} \end{array}$$

em que n_{ij} é o número de casos na categoria i do item 1 e da categoria j do item 2. Suponha que esses itens são de variáveis normalmente distribuídas, Z_1^* e Z_2^*. Assume-se que a sua distribuição conjunta é normal bivariada com correlação ρ. A correlação policórica é a correlação ρ da distribuição normal bivariada $N(0, 0, 1, 1, \rho)$ (F.1) das variáveis latentes Z_1^* e Z_2^*. Se $m_1 = m_2 = 2$, então a correlação é tetracórica.

$$P[X = 1, Y = j] = p_{ij} = \int_{a_i - 1}^{a_1} \int_{b_j - 1}^{b_1} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp^{-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)} \tag{F.1}$$

e pode ser estimada maximizando a função de máxima verossimilhança da distribuição multinomial:

$$\ln L = \sum_{i=1}^{m1} \sum_{j=1}^{m2} n_{ij} \log p_{ij}. \tag{F.2}$$

A correlação policórica é utilizada quando as variáveis são contínuas, linearmente relacionadas e divididas em categorias.

Apêndice G

Algoritmo R

```
# Dois estágios
rm(list = ls())
library(readstata13)

library(Matrix); library(MASS); library(numDeriv);
library(xtable); library(psych);

Odata <- read.dta13('Base_5_ano_filtrada.dta')
Odata0 <- na.omit(Odata)

end.form <- inse_3 ~ ln_ndoc_pa + as.factor(in_biblioteca_sala_leitura) +
ln_nu_computador_pa + ln_tempo_5 + as.factor(urbano) +
ln_historico_escolar_5 + ln_atencao_pais_5 + ln_acesso_cultura_5 +
ln_ded_extra_classe_5 + ln_afd_fun12 + ln_ied_fun123 + as.factor(icg_3) +
ln_mird + uf1  + uf2 +  uf3 +  uf4+ uf5 + uf6 + uf7 + uf8 + uf9+ uf10 + uf11
+ uf12 + uf13 + uf14 + uf15 + uf16 + uf17 + uf18 + uf19 + uf20 + uf21 + uf22 +
uf23 + uf24 + uf25 + uf26

{
Reglogistica <- function(end.form, data = sys.parent()){

# Variaveis endogenas
```

```
End <- end.form
m <- model.frame(End, data)
xend <- model.matrix(End, m)
yend <- model.response(m, "numeric")

yres <- log(yend + 0.01 /(1.01 - yend))

    dados0 <- data.frame(data, yres)

    form.two <- update(end.form, yres ~ . )

    result <- lm( form.two,data = dados0 )

    parms <- result[1]$`coefficients`

    # Passo 1 -  variaveis endogenas
    k6 <- dim(xend)[2]

    lli <- function(z){
      betas <- z[1:k6]
      eta_betas <- as.vector(xend %*% betas)
      y.est <- (1/(1 + exp(-(eta_betas))))

      ll <- sum(yend*log(y.est) + (1 - yend)*log(1 - y.est))

      return(-ll)
    }

    gg <- function(z){
      betas <- z[1:k6]
      eta_betas <- as.vector(xend %*% betas)
      y.est <- (1/(1 + exp(-(eta_betas))))
```

```
      dif_y <- yend - y.est
      g <- t(dif_y) %*% xend
      return(-g)
    }

    #####
    est1_frac <- optim(par = parms, fn = lli, gr = gg,
                       hessian = TRUE,
                       method = "BFGS",
                       control = list(trace=TRUE, REPORT=1, fnscale = 1))
    b <- est1_frac$par
    V1 <- solve(est1_frac$hessian)
    ep <- sqrt(diag(V1))
    estZ <- lapply(length(b), function(x) b/ep)
    p_valor <- lapply(length(b), function(x) (1 - pnorm(abs(b/ep)))*2)
    ics <- lapply(length(b), function(x) cbind(b - qnorm(.975) * ep,
    b + qnorm(.975) * ep))

    result <- data.frame("Coef" = b, "Std.Err" = ep, "z" = estZ[[1]],
    "p-value" = p_valor[[1]], "LI" = ics[[1]][,1], "LS" = ics[[1]][,2])
    result0 <- list("Reg" = result, "COV_END" = V1)
    return(result0)
  }
}

resultado <- Reglogistica(end.form = end.form, data = Odata0)
resultado

tabela_log<-xtable(resultado$Reg ,digits=3)
display(tabela_log)[c(1,4:5)]<-"e"
```

```
End <- end.form
m <- model.frame(End, Odata0)
xend <- model.matrix(End, m)

eta0 <- xend %*% resultado[[1]]$Coef

y.endg <- 1 /  (1 + exp(- eta0))
# V1 <- resultado$COV_END
dados0 <- data.frame(Odata0, y.endg)

fr.form <- ln_media_nota_5 ~ ln_ndoc_pa + as.factor(in_biblioteca_sala_leitura) +
ln_nu_computador_pa + ln_tempo_5
s2u.form <- ~ y.endg + ln_historico_escolar_5 + ln_atencao_pais_5 +
ln_acesso_cultura_5 + ln_ded_extra_classe_5 + ln_afd_fun12 +
ln_ied_fun123 + as.factor(icg_3) + ln_mird
s2w.form <- ~ uf1 + uf2 + uf3 + uf4 + uf5 + uf6 + uf7 + uf8 + uf9 + uf10 + uf11
+ uf12 + uf13 + uf14 + uf15 + uf16 + uf17 + uf18 + uf19 + uf20 + uf21 + uf22
+ uf23 + uf24 + uf25 + uf26 + as.factor(urbano)

data <- dados0

# pamrsEnd <- resultado[[1]]$Coef

{

  SF.half2S <- function(fr.form, end.form, s2u.form, s2w.form, pamrsEnd, V1,
  data = sys.parent()){
    {# Fronteira
      frontier <- fr.form
      mf <- model.frame(frontier, data)
      X_fr <- model.matrix(frontier, mf)
```

```
  Y_fr <- model.response(mf, "numeric")

  # Sigma_u e Sigma_w
  Sigma_u <- s2u.form
  sig_u <- model.frame(Sigma_u, data)
  sigma_u <- model.matrix(Sigma_u, sig_u)

  Sigma_w <- s2w.form
  sig_w <- model.frame(Sigma_w, data)
  sigma_w <- model.matrix(Sigma_w, sig_w)

  # Endogenidade
  End <- end.form
  m <- model.frame(End, data)
  xend <- model.matrix(End, m)
  yend <- model.response(m, "numeric")

  k6 <- dim(xend)[2]

  # Valores iniciais
  fr <- lm(frontier, data = data)
  u.p <- c(1, rep(0, ncol(sigma_u) - 1)); names(u.p) <- colnames(sigma_u)
  w.p <- c(1, rep(0, ncol(sigma_w) - 1)); names(w.p) <- colnames(sigma_w)

  z <- c(coef(fr), u.p, w.p)

  kk <- length(z)
  k1 <- length(coef(fr))
  k2 <- k1 + length(u.p)
  k3 <- k2 + length(w.p)
}
{
```

```
ll <- function(z){
  x_p <- z[1:k1]
  u_p <- z[c(k1 + 1):k2]
  w_p <- z[c(k2 + 1):k3]

  Y <- X_fr %*% x_p
  U <- sigma_u %*% u_p
  W <- sigma_w %*% w_p

  Sigma_u <- exp(U)
  Sigma_w <- exp(W)
  lambda <- sqrt(Sigma_u/Sigma_w)
  sigma <- Sigma_u + Sigma_w

  e <- Y_fr - Y

  ly.x.half <- 0.5 * log(2 / pi) - 0.5 * log(sigma) +
  pnorm(-lambda * e / sqrt(sigma), log = TRUE) - 0.5 * e^2 / sigma

  val <- sum(ly.x.half)

  return(-val)
}
G <- function(z){
  x_p <- z[1:k1]
  u_p <- z[c(k1 + 1):k2]
  w_p <- z[c(k2 + 1):k3]

  Y <- X_fr %*% x_p
  U <- sigma_u %*% u_p
  W <- sigma_w %*% w_p
```

```
    Sigma_u <- exp(U)
    Sigma_w <- exp(W)
    lambda <- sqrt(Sigma_u/Sigma_w)
    sigma <- Sigma_u + Sigma_w

    e <- Y_fr - Y

    zz <- - e * lambda / sqrt(sigma)
    dz <- pmax(dnorm(zz), 9.88131291682493e-324)
    pz <- pmax(pnorm(zz), 9.88131291682493e-324)
    fdp <- dz; cdf <- pz; fdp_cdf <- fdp / cdf
    valor <- as.vector(e / sigma + lambda / sqrt(sigma) * fdp_cdf)

    g.b <- t(X_fr) %*% valor
    g.u <- t(sigma_u) %*% as.vector((0.5 / sigma *
    (e^2 / sigma - e / (lambda * sqrt(sigma)) * fdp_cdf - 1)) * Sigma_u)
    g.w <- t(sigma_w) %*% as.vector((0.5 / sigma *
    (e^2 / sigma + e * lambda * (2 + lambda^2) * fdp_cdf / sqrt(sigma) - 1))
    * Sigma_w)

    g <- c(g.b, g.u, g.w)

    return(-g)
  }
}

est <- optim(par = z, fn = ll, gr = G, hessian = TRUE, method = "BFGS",
             control = list(fnscale = 1, trace = TRUE, REPORT = 1,
             maxit = 1000))
#round(data.frame("numerico" = grad(ll,est$par), "analitico" = G(est$par)), 4)
{
  gg0 <- function(z){
```

```
  betas <- z[1:k6]
  eta_betas <- as.vector(xend %*% betas)
  y.est <- (1/(1 + exp(-(eta_betas))))

  dif_y <- yend - y.est
  g <- xend * as.vector(dif_y)

  return(g)
}
G11 <- gg0(pamrsEnd)



b <- est$par
yest <- X_fr %*% b[1:k1]
s2u <- exp(sigma_u %*% b[c(k1 + 1):k2])
s2w <- exp(sigma_w %*% b[c(k2 + 1):k3])
sigma <- s2u + s2w
lambda <- sqrt(s2u/s2w)
erro <- Y_fr - yest

G2 <- function(z){
  x_p <- z[1:k1]
  u_p <- z[c(k1 + 1):k2]
  w_p <- z[c(k2 + 1):k3]



  Y <- X_fr %*% x_p
  U <- sigma_u %*% u_p
  W <- sigma_w %*% w_p

  Sigma_u <- exp(U)
  Sigma_w <- exp(W)
```

```
  lambda <- sqrt(Sigma_u / Sigma_w)
  sigma <- Sigma_u + Sigma_w

  e <- Y_fr - Y

  zz <- - e * lambda / sqrt(sigma)
  dz <- pmax(dnorm(zz), 9.88131291682493e-324)
  pz <- pmax(pnorm(zz), 9.88131291682493e-324)
  fdp <- dz; cdf <- pz; fdp_cdf <- fdp / cdf
  valor <- as.vector(e / sigma + lambda / sqrt(sigma) * fdp_cdf)

  g.b <- X_fr * valor
  g.u <- sigma_u * as.vector((0.5 / sigma *
  (e ^ 2 / sigma - e / (lambda * sqrt(sigma)) * fdp_cdf - 1)) * Sigma_u)
  g.w <- sigma_w * as.vector((0.5 / sigma *
  (e ^ 2 / sigma + e * lambda * (2 + lambda ^ 2) *
  fdp_cdf / sqrt(sigma) - 1)) * Sigma_w)

  g <- cbind(g.b, g.u, g.w)

  return(g)
}
G22 <- G2(b)

g21 <- function(z){
  x_p <- z[1:k1]
  u_p <- z[c(k1 + 1):k2]
  w_p <- z[c(k2 + 1):k3]


  Y <- X_fr %*% x_p
  U <- sigma_u %*% u_p
```

```
  W <- sigma_w %*% w_p

  Sigma_u <- exp(U)
  Sigma_w <- exp(W)
  lambda <- sqrt(Sigma_u / Sigma_w)
  sigma <- Sigma_u + Sigma_w

  e <- Y_fr - Y

  zz <- - e * lambda / sqrt(sigma)
  dz <- pmax(dnorm(zz), 9.88131291682493e-324)
  pz <- pmax(pnorm(zz), 9.88131291682493e-324)
  fdp <- dz; cdf <- pz; fdp_cdf <- fdp / cdf
  valor <- as.vector(e / sigma + lambda / sqrt(sigma) * fdp_cdf)

  g.u <- xend * as.vector((0.5 / sigma * (e ^ 2 / sigma - e /
  (lambda * sqrt(sigma)) * fdp_cdf - 1)) * Sigma_u * u_p[2] *
  sigma_u[,2]*(1 - sigma_u[,2])  )

  return(g.u)
}
G21 <- g21(b)

V1 ; V2 <- ginv(est$hessian); C <- t(G22) %*% G21; R <- t(G22) %*% G11
V2_adj <- V2 + V2 %*% ((C %*% V1 %*% t(C)) - (R %*% V1 %*% t(C))
- (C %*% V1 %*% t(R))) %*% V2

ep <- sqrt(diag(V2_adj))
estZ <- lapply(length(b), function(x) b/ep)
p_valor <- lapply(length(b), function(x) (1 - pnorm(abs(b/ep)))*2)
ics <- lapply(length(b), function(x) cbind(b - qnorm(.975) * ep,
b + qnorm(.975) * ep))
```

```
      mu.mod <- - erro * s2u / sigma
      s.mod <- sqrt(s2w * s2u / sigma)
      uf <- mu.mod + s.mod * ((dnorm(-mu.mod / s.mod))
      / (pnorm(mu.mod / s.mod)))
      ef <- ((pnorm(-s.mod + mu.mod / s.mod)) / (pnorm(mu.mod / s.mod)))
      * exp(- mu.mod + 0.5 * s.mod^2)

      result <- data.frame("Coef" = cbind(est$par), "Std.Err" = ep,
      "z" = estZ[[1]],"P-value" = p_valor[[1]], "LI" = ics[[1]][,1],
      "LS" = ics[[1]][,2])

    }
    lista <- list("eff" = ef, "erro" = erro, "table" = result,
    'V2-semcorrecao' = V2,"summary.ef" = summary(ef),
    "sd.ef" = sd(ef, na.rm = TRUE), "V2_corrigida" = V2_adj)
    return(lista)
  }

}
mod.h2 <- SF.half2S(fr.form = ln_media_nota_5 ~ ln_ndoc_pa +
as.factor(in_biblioteca_sala_leitura) + ln_nu_computador_pa + ln_tempo_5,
s2u.form = ~ y.endg + ln_historico_escolar_5 + ln_atencao_pais_5 +
ln_acesso_cultura_5 + ln_ded_extra_classe_5 + ln_afd_fun12 + ln_ied_fun123 +
as.factor(icg_3) + ln_mird,
s2w.form = ~ uf1  + uf2 +  uf3 +  uf4+  uf5 + uf6 + uf7 + uf8 + uf9 +
uf10 + uf11  +uf12+  uf13+  uf14+  uf15  +uf16 + uf17  +uf18 + uf19 +
uf20 + uf21+  uf22+  uf23+  uf24+  uf25+  uf26 + as.factor(urbano),
end.form = inse_3 ~ln_ndoc_pa + as.factor(in_biblioteca_sala_leitura) +
ln_nu_computador_pa + ln_tempo_5 + as.factor(urbano) +
ln_historico_escolar_5 + ln_atencao_pais_5 + ln_acesso_cultura_5 +
ln_ded_extra_classe_5 +ln_afd_fun12 + ln_ied_fun123 + as.factor(icg_3) +
```

```
ln_mird + uf1  + uf2 +  uf3 +  uf4+  uf5 + uf6 + uf7 + uf8 + uf9+ uf10 +
uf11  +uf12+  uf13+  uf14+  uf15  +uf16 + uf17  +uf18 + uf19+  uf20 +
uf21+  uf22+  uf23+  uf24+  uf25+  uf26,
pamrsEnd = resultado[[1]]$Coef, V1 = resultado$COV_END,
data = dados0)
```