ipea Instituto de Pesquisa Econômica Aplicada

Título do capítulo	CAPÍTULO 6 **A OPERACIONALIZAÇÃO DA AVALIAÇÃO DE EFICIÊNCIA ECONÔMICA: AS PROPRIEDADES DAS MEDIDAS E OS PRINCIPAIS MÉTODOS DE CÁLCULO**
Autores (as)	Alexandre Marinho
DOI	

Título do livro	SUS: AVALIAÇÃO DA EFICIÊNCIA DO GASTO PÚBLICO EM SAÚDE
Organizadores (as)	Carlos Octávio Ocké-Reis
Volume	
Série	
Cidade	
Editora	Instituto de Pesquisa Econômica Aplicada (Ipea)
Ano	2023
Edição	1ª
ISBN	
DOI	

A operacionalização da avaliação de eficiência econômica: as propriedades das medidas e os principais métodos de cálculo

Alexandre Marinho

Introdução

No capítulo 2, vimos que nem todos os tipos de indicadores úteis para avaliar o desempenho de uma unidade produtiva são indicadores de eficiência. À guisa de ilustração, o conceito de eficiência de PK é muito utilizado por economistas, mas é pouco difundido fora de ciência econômica. Outros conceitos e noções de apoio também podem ser válidos.

Entretanto, para que um conceito origine uma medida de eficiência razoável, algumas características mínimas devem ser observadas, conforme discutido em Bogetoft e Otto[1], que descrevemos, com aconselhamentos de ordem prática, neste capítulo, no qual vamos expor, de forma sucinta, os principais métodos para mensuração de eficiência técnica, alocativa e de escala, com as suas principais características. Sempre que for necessário, forneceremos referências ao leitor interessado em se aprofundar em detalhes mais técnicos.

Propriedades Gerais

Vejamos abaixo as propriedades gerais de uma medida de eficiência robusta:

1. Flexibilidade: para ter interesse geral, um conceito de eficiência deve ser aplicável a uma ampla classe de tecnologias. Não é razoável que se elabore uma medida de

eficiência para cada problema particular. O melhor guia para evitar esse problema é a literatura sobre o assunto, disponível em bons periódicos científicos e em livros publicados por editoras com reputação.

2. Métrica: uma medida de eficiência, basicamente, mapeia um plano de produção e uma tecnologia no conjunto dos números reais. Isso significa que a medida de eficiência deve ser um número real. Normalmente, um escore de eficiência é um número puro ou um percentual. Deve-se trabalhar, sempre que possível, com números positivos para insumos e produtos, de modo que os escores também sejam sempre positivos, embora alguns *softwares*[I] permitam trabalhar com insumos e produtos negativos (mas não com escores negativos). Números negativos não são intuitivos em se tratando de teoria da produção, que lida com quantidades de bens físicos. Contudo, em alguns contextos específicos, números negativos podem ocorrer, por exemplo, em virtude de transformações em variáveis, ou para retratar prejuízos financeiros e contábeis. Também é aconselhável trabalhar com escores no intervalo [0, 1] ou [0, 100%]. Nesse intervalo, uma DMU com eficiência igual à unidade ou 100% é dita eficiente, e quanto maior o escore, maior será a eficiência e vice-versa. Entretanto, como vimos na discussão, nas funções distância, em um modelo de insumo orientado de Shepard e em um modelo produto orientado de Farrell, os escores de eficiência estão no intervalo [1, infinito], e quanto menor o escore, melhor a avaliação. Nesses casos, como não há limite superior para o escore, a medida perde em intuição, pois não se sabe, *a priori*, se um determinado escore pode ser considerado relativamente alto ou baixo. Por isso, nessas situações, é comum inverter os escores, para que eles fiquem no intervalo [0, 1].

3. Comensurabilidade ou invariância para permutações e reescalonamento: não importa a ordem de apresentação de insumos e de produtos. Também não interessa se reescalonamos todos os diferentes insumos e produtos. Por exemplo, não importa se medimos insumos e produtos em quilogramas ou toneladas, em litros ou hectolitros, desde que eles sejam medidos nas mesmas unidades em todas as DMU. Então, não é possível medir a receita de um hospital em dólares e comparar com a receita de outro hospital medida em reais.

4. Indicação: somente os pontos eficientes sob PK são plenamente eficientes. A medida de Farrell não tem essa propriedade, e a medida de PK tem, obviamente. Trata-se de uma propriedade mais técnica, mas cuja desconsideração pode acarretar sérias

I O ***software Benchmarking*** é um exemplo. Ver Bogetoft e Otto[1].

consequências na avaliação. Se um modelo não computa as folgas nas isoquantas (trechos horizontais ou verticais, em que ocorre EPf, que vimos na definição de ET (capítulo 2); pode-se não perceber a possibilidade de economia de insumos ou de expansão de produtos. Isso ocorreria porque a medida de eficiência adotada é apenas radial, ou seja, considera apenas a possibilidade de aumento equiproporcional em produtos, ou redução equiproporcional em insumos, mas não leva em conta as folgas isoladas em alguns dos produtos ou os excessos isolados em alguns dos insumos. Ambas as folgas são chamadas de *slacks* na língua inglesa.

5. Homogeneidade de grau 1: se multiplicamos os insumos por uma constante positiva, dividimos a eficiência pela mesma constante. Se multiplicamos os produtos por uma mesma constante positiva, a eficiência dobra. Note-se que não estamos falando de retornos de escala. Não estamos avaliando o efeito, nos produtos, da multiplicação ou divisão dos insumos por uma constante. Estamos falando no efeito sobre os escores de eficiência. Então, se um hospital tem eficiência igual a 100% e dobramos o seu uso de insumos, mantido o nível de produtos, a sua eficiência deverá será agora de 50%.

6. Monotonicidade: se aumentamos o consumo de pelo menos um insumo, sem aumento de algum produto, reduzimos a eficiência e vice-versa. Fenômeno análogo ocorre na redução de algum produto sem a redução de algum insumo reciprocamente. Essa propriedade não existe na eficiência de Farrell, em que os aumentos e as reduções que afetam a eficiência devem ser equiproporcionais, mas existe em PK.

7. Continuidade: se variarmos o consumo de insumos, ou a produção de produtos, de modo parcimonioso, o escore de eficiência deve variar continuamente. A medida de eficiência não deve ter sensibilidade extrema, pois não desejamos que pequenos erros nos dados tenham impactos dramáticos na eficiência. Existe em Farrell e em PK.

Avaliação de eficiência e incerteza

As incertezas são grandes em saúde, podem surgir em diferentes circunstâncias e, de modo esquemático não exaustivo, podem abranger:

1. o grau de custo-efetividade da prevenção;

2. a precisão de diagnósticos;

3. as eficácias de prescrições alternativas;

4. os possíveis resultados esperados (prognósticos) do curso da doença com e/ou sem tratamento;

5. a magnitude dos custos, embora a presença dos custos seja certa. Nessa certeza, algo singular ocorre, porque não existe saúde grátis. Mesmo em caso de morte – evitável ou não –, existe uma conta a ser paga por aqueles que ficam vivos. Ou então, o tratamento ou já foi financiado *ex ante* por alguém que pode ser o próprio paciente, pagando do próprio bolso no momento da entrada no serviço (*out-of-pocket*), ou via planos e seguros de saúde, ou via impostos.

Os custos e benefícios dos cuidados em saúde têm valores incertos, *a priori*, e, quase sempre, difíceis de observar e medir (ao menos em grande escala ou em largos períodos de observação). Então, as atividades de avaliação, ao criar referências sobre os desempenhos em uma amostra, agregam informação ao processo decisório em saúde, reduzem as incertezas e, portanto, podem melhorar esse processo.

Como a avaliação de eficiência tem como condição necessária algum grau de monitoramento, e de padronização de indicadores na amostra analisada, ela é um importante instrumento de coordenação e de regulação setorial (*monitoring principle*). Por exemplo, um hospital pode ter sido muito mal (ou muito bem) avaliado, porque não coletou corretamente os seus dados, ou porque sua tecnologia é muito diferente dos demais. Ao observar as melhores práticas (*benchmarks*) reveladas pela avaliação, podem ser visualizadas melhores tecnologias, processos produtivos mais seguros, estratégias mais adequadas, e maior grau de *compliance*. Além disso, a avaliação pode elevar o moral de seu *staff* ao mostrar que é possível melhorar o desempenho, revelando metas e processos produtivos factíveis, e reduzir a presença de expectativas excessivamente pessimistas ou otimistas.

Por outro lado, a descoberta de desempenhos abaixo do esperado em uma unidade pode suscitar a melhoria dos processos produtivos dessa unidade. No entanto, também pode revelar a precariedade de coleta de dados necessários para avaliação na unidade mal avaliada, que não valoriza o processo avaliativo, eventualmente porque ele nunca foi feito de modo sério, ou correto, ou sistemático. Surge, inclusive, a possibilidade de que haja necessidade de revisão do processo avaliativo, que não contemplava corretamente determinadas peculiaridades na estrutura ou nas missões de determinadas organizações. Imagine-se o que pode acontecer se uma universidade de grande porte, que tenha um ou mais hospitais universitários, for comparada, sem maiores cuidados, com uma faculdade de medicina que não tenha um hospital próprio. Esse é um problema comum que ocorre quando, por exemplo, surgem, na mídia, comparações entre universidades

públicas e privadas, ou entre universidades brasileiras e estrangeiras, baseadas apenas na relação entre a quantidade de professores e a quantidade de alunos (a famigerada relação professor/aluno).

Taxonomia dos métodos disponíveis de avaliação de eficiência

A maioria dos modelos de avaliação eficiência econômica se enquadra, com maior ou menor pertinência, na seguinte taxonomia, referente às quantidades de parâmetros que serão estimados:

1. Métodos paramétricos: são definidos *ex ante*, ou seja, a função matemática de transformação de insumos em resultados é totalmente definida, com exceção de um conjunto finito de parâmetros que são estimados a partir dos dados. Exemplos: Análise de Fronteiras Estocásticas (*Stochastic Frontier Analysis* – SFA) e Mínimos Quadrados Ordinários Corrigidos (*Corrected Ordinary Least Squares* – COLS).

2. Métodos não paramétricos: são muito menos restritivos. Apenas um amplo conjunto de classes de funções é estabelecida *a priori* (por exemplo: funções de produção convexas, crescentes nos insumos). As classes são tão amplas que não permitem o estabelecimento de um conjunto limitados de parâmetros a serem estimados. Por isso, são denominadas de não paramétricas. Exemplos; *Free Disposal Hull* (FDH) e Análise Envoltória de Dados (*Data Envelopment Analysis* – DEA).

Uma segunda taxonomia, que não exclui a anterior, tem a ver com a influência das aleatoriedades nos desempenhos observados. Temos basicamente as seguintes categorias:

1. Métodos estocásticos: admitem que as observações (os dados) individuais podem ser afetadas por ruídos (perturbações ou choques) aleatórios (fortuitos) causados por diversos fatores, como influências externas, erros de medidas etc. O desempenho estimado dos agentes é extirpado desses ruídos. Exemplo: SFA e Análise Envoltória de Dados Estocástica (*Stochastic Data Envelopment Analysis* – SDEA).

2. Métodos determinísticos: nesses casos, o ruído aleatório não é considerado, e toda a variação nos dados é relevante para a determinação do desempenho. Há um ponto importante: a verdadeira fronteira de eficiência não é observada. Então, os estimadores gerados por esses métodos são originados por processos de geração de dados (*Data Generating Process* – DGP) desconhecidos. Consequentemente, os escores de eficiência gerados nesses métodos são estimadores com distribuição de probabilidades desconhecidas, o que não permitiria a produção de intervalos de confiança nem testes de hipóteses. Ademais, como a eficiência observada na amostra não pode ser maior do que a eficiência real,

os escores são viesados em favor das DMU (benevolentes). Adicionalmente, quando os escores são relativos, eles são correlacionados. Em resumo: há que se ter muito cuidado no trabalho com algumas dessas metodologias, principalmente se houver a intenção de combiná-las com métodos econométricos para, por exemplo, tentar encontrar determinantes da eficiência que estejam fora do controle dos administradores ou gestores das DMU[2,3]. Esses problemas são particularmente importantes em DEA, que é um método de programação matemática com utilização crescente em saúde[4,5], ao qual retornaremos.

O cálculo dos escores de eficiência

Independentemente do método escolhido, a maioria dos métodos de estimação de eficiência têm, ao menos, uma característica em comum: a geração de um *ranking* de desempenho entre as DMU, por meio do cálculo de escores de eficiência. A construção desse *ranking* é sempre possível, mesmo na presença de múltiplos indicadores de insumos e produtos que poderiam embaralhar a ordenação das DMU, por um motivo simples: os métodos calculam pesos ótimos que permitem ponderar (agregar) os diferentes indicadores. Esses pesos são obtidos fazendo alguma espécie de comparação entre as DMU e escolhendo, como referências para a construção de uma fronteira de eficiência, as unidades que atenderem ao critério de eficiência de PK, ou de Farrell, o de Shephard, que apresentamos no capítulo 2, ou de outros alternativos (por exemplo: eficiência direcional, ou eficiência hiperbólica)[II]. Sem detalhar as tecnicalidades, existem basicamente duas maneiras de gerar os pesos ótimos, que descrevemos brevemente a seguir:

1. Modelos de regressão[6]: nesse caso, os custos (ou insumos) e produtos são incorporados em modelos de regressão[III]. Os coeficientes angulares da regressão (os *betas*) permitem obter uma medida agregada de relação entre insumos (ou entre os custos) e os produtos, gerando uma fronteira de eficiência advinda de uma função de produção ou, alternativamente, gerando uma função custo, especificadas *ex ante* (por exemplo: Cobb-Douglas, Translogarítmica). Alguma medida dos desvios (distâncias) das DMU em relação à fronteira (considerando ou não os erros aleatórios) gera um *ranking* de desempenho. Os exemplos mais conspícuos são os modelos de SFA e os modelos COLS.

2. Modelos de programação matemática[7]: nesse caso, são construídas combinações lineares, convexas ou não, com pesos ótimos, dos vetores de insumos (ou de custos)

II Ver Bogetoft e Otto[1].

III Para uma introdução aos modelos de regressão, consultar Maddala[6].

e dos vetores de produtos das diferentes alternativas[IV]. Assim, é possível detectar quais as DMU que formam a fronteira de eficiência e quais estão aquém da fronteira. Dito de outro modo, os modelos permitem agregar os insumos de todas as DMU, e os seus produtos, com os melhores pesos possíveis (pesos ótimos) respeitada a restrição de que nenhuma DMU pode estar além da fronteira de eficiência. Os modelos mais conhecidos são DEA; FDH; e os Índices de Malmquist (*Malmquist Index* – MI)[V].

Além de gerar um ranking com os valores das eficiências das DMU, os modelos de eficiência listados acima, resguardadas as especificidades, permitem calcular, para cada elemento DMU da amostra (organizações e programas de saúde, ações em saúde etc.), os valores ótimos de cada um dos componentes dos custos (ou insumos) e benefícios (resultados ou produtos) quantitativos utilizados. Também podem permitir especificar a natureza dos retornos de escala da amostra, bem como indicar quais são as referências (*benchmarks*) da amostra para cada um dos seus elementos DMU. Por exemplo: se um hospital avaliado como ineficiente utilizou médicos, enfermeiros, leitos e medicamentos para produzir exames, consultas e cirurgias, deverá ser possível, a partir de alguns modelos, obter os valores ótimos de cada uma dessas variáveis que tornaria o referido hospital eficiente. Talvez mais interessante é o fato de que, agregando os resultados individuais, deverá ser possível calcular os valores ótimos da amostra, que pode ser um conglomerado de unidades de saúde, um município, uma região de saúde, um sistema de saúde, um estado, um país etc.

Os modelos mais usados de avaliação de eficiência em saúde

Dois modelos se destacam na avaliação de eficiência econômica de entidades em saúde: DEA e SFA. Ambos os métodos são bastante flexíveis e possuem muitas variações técnicas e possibilidades de combinações com outros métodos. Como a DEA é, em princípio, determinística e não paramétrica, e a SFA é estocástica e paramétrica, elas cobrem uma ampla gama de possibilidades metodológicas que apresentamos nas taxonomias descritas na seção precedentes. Ambos os métodos são chamados de 'modelos de fronteira de eficiência', porque uma fronteira é o lugar geométrico das DMU eficientes: estas, operando abaixo da fronteira de produção ou acima de uma fronteira de custos, são ditas ineficientes.

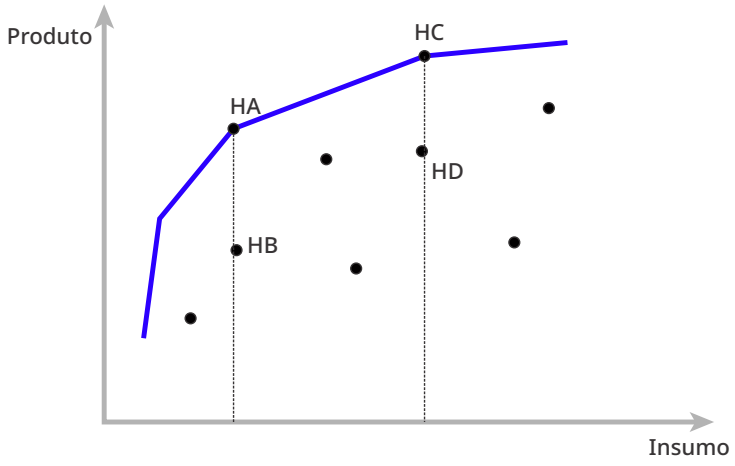
IV Para uma introdução aos modelos de programação matemática, consultar Hillier e Lieberman[7]

V O ***software*** livre R (https://www.r-project.org/) abriga vários programas gratuitos que executam todos esses modelos. São inúmeros os softwares comerciais disponíveis para fazer essas análises.

Na figura 1, temos uma fronteira de produção. O hospital HA gasta a mesma quantidade de insumos que o hospital HB, mas gera uma quantidade maior de produto. Diz-se que o hospital HA domina o hospital HB. O hospital HC gasta a mesma quantidade de insumos que o hospital HD, mas gera uma quantidade maior de produto. Diz-se que o hospital HC domina o hospital HD. Os hospitais HA e HC são eficientes, e os hospitais HB e HD são ineficientes. Vale assinalar que nenhuma DMU dominada por qualquer outra DMU pode ser eficiente. A fronteira de eficiência de produto é obtida conectando os pontos eficientes por meio de segmentos de reta, como o segmento HA-HC. Todos os pontos abaixo da fronteira são ineficientes. Contudo, a comparação entre HB e HD não é trivial. HB não domina nem é dominada por HD, pois HB produz menos do que HD, mas também usa menos insumos. O modelo de cálculo de escores de eficiência vai gerar o *ranking* entre todas as DMU, inclusive HB e HD.

A introdução de convexidade na fronteira facilita visualizar as firmas ineficientes[VI]. De fato, a convexidade conecta pontos isolados (DMU) por meio de segmentos de retas como o segmento HA-HC. A convexidade permite comparar até mesmo eventuais firmas inexistentes, ou combinações de firmas existentes. Nenhuma DMU, além de HA e HC, foi observada no segmento HA-HC na realidade. Mas qualquer DMU, existente ou não, que fosse localizada abaixo desse segmento, ou dos demais segmentos de reta da fronteira, seria ineficiente.

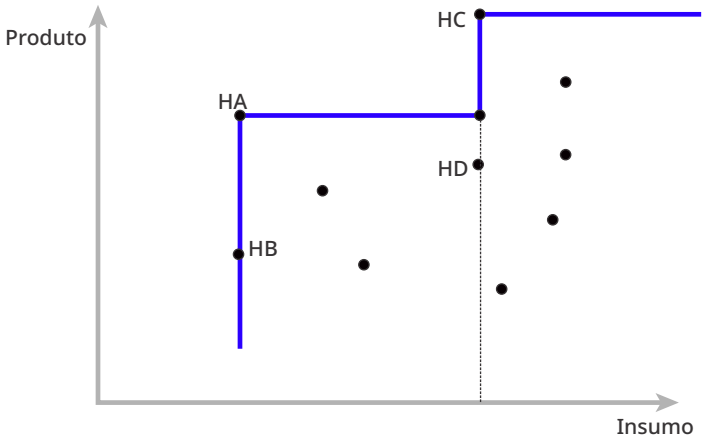
Figura 1. Fronteira de produção



VI Convexidade: se duas combinações de insumos e de produtos são factíveis, então qualquer combinação convexa dessas combinações também será factível. Seja T a tecnologia. Se $(x_j, y_j) \in T$, $\lambda_j \geq 0$, $\sum_{j}^{n}=1 \lambda_j=1$, $j=1...n$. Então: $(\sum_{j}^{n}=1 \lambda_j x_j, \sum_{j}^{n}=1 \lambda_j y_j) \in T$.

Entretanto, se o analista não quer usar um modelo que admita convexidade, e deseja construir uma fronteira apenas com DMU efetivamente existentes, uma opção é o método FDH, cuja fronteira de eficiência não é convexa[VII], representado na figura 2 a seguir. O modelo FDH também está disponível em grande parte dos *softwares* disponíveis, como o já citado *Benchmarking*. De fato, já existem testes disponíveis para avaliar a convexidade e a natureza dos retornos de escala do conjunto de produção[8].

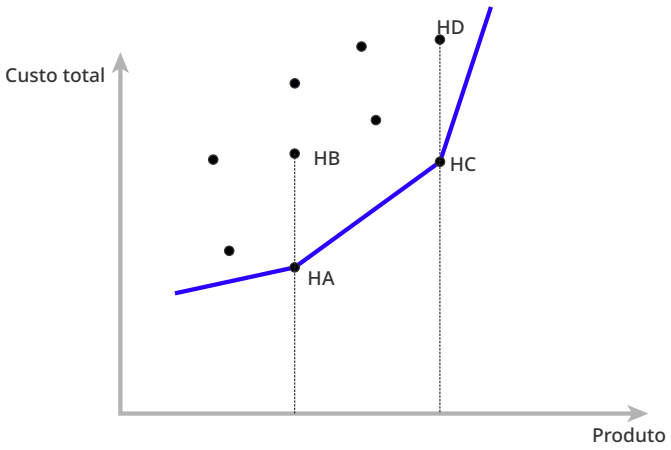
Figura 2. *Free Disposal Hull*



Na figura 3, temos uma fronteira de custos que considera os custos totais de produção de cada hospital. O hospital HA gasta o mesmo que o hospital HB, mas gera maior quantidade de produto. O hospital HC gasta o mesmo que o hospital HD, mas gera maior quantidade de produto. Os hospitais HA e HC são eficientes, e os hospitais HB e HD são ineficientes. Novamente, como na fronteira da produção, a comparação entre os hospitais HB e HD não é trivial, e depende dos cálculos dos escores de cada um deles, no modelo de eficiência que estiver sendo empregado. A fronteira de eficiência de custos é obtida conectando os pontos eficientes por meio de segmentos de reta. Todos os pontos acima da fronteira são ineficientes.

VII Um conjunto é dito convexo quando os pontos sobre um segmento de reta unindo quaisquer dois pontos pertencentes ao conjunto também pertencem ao conjunto.

Figura 3. Fronteira de custo



Análise Envoltória de Dados (DEA)

A DEA é um método não paramétrico, utilizado para avaliar a eficiência de um conjunto com N unidades produtivas chamadas de DMU, que transformam múltiplos insumos em múltiplos produtos. A DEA tem as suas raízes no trabalho de Farrell[9], que introduziu o conceito de isoquanta[VIII]. Posteriormente, o trabalho de Charnes, Coper e Rhodes[10] utilizou um modelo de programação matemática para operacionalizar o método.

No caso mais geral, uma DMU utiliza múltiplos insumos descritos no vetor[IX] $X=(x_1,...,x_s)$ para produzir múltiplos produtos descritos no vetor $Y=(y_1,...,y_m)$; e o seu escore de eficiência será definido pelo seguinte quociente ponderado entre um "produto virtual" e um "insumo virtual":

$$\text{Eficiência}=(u_1y_1+...+u_my_m)/(v_1x_1+.....v_sx_s).$$

Os insumos $X=(x_1,...,x_s)$ e os produtos $Y=(y_1,...,y_m)$ são os dados observados em cada DMU. No cálculo da eficiência, são usados um vetor de pesos não negativos, associados aos insumos, dado por $V=(v_1,...,v_s)$, e um vetor de pesos não negativos, associados aos produtos, dado por $U=(u_1,...,u_m)$. Esse escore de eficiência será calculado, para cada DMU, por um problema de programação matemática que escolherá os vetores de 'pesos ótimos que maximizam a eficiência', levando em conta a restrição, nos modelos básicos mais simples, de que esse quociente não seja maior do que a unidade. Uma DMU eficiente terá

VIII Ver capítulo 2.

IX Ao leitor não familiarizado com matemática aplicada em economia, esclarecemos que, no presente contexto, um vetor é um conjunto (uma lista) de diferentes insumos, ou produtos, apresentando as respectivas quantidades. Exemplo: 3 enfermeiros, 1 médico, 1 leito; ou 10 cirurgias, 3 óbitos etc.

escore igual a unidade ou 100%. Cabe ao avaliador ter sensibilidade para não tachar de ineficientes DMU com escores muito próximos da unidade, dada a carga pejorativa desse termo em seu uso coloquial. É preciso ter muito cuidado com o uso do termo ineficiente em economia, que não significa má qualidade, ruim, descartável etc. – assim como ser eficiente não significa ter alta qualidade, ser bom, ou qualificações similares. O conceito de eficiência em economia tem significados precisos e limitados, como estamos vendo. Ademais, alguns modelos de DEA modernos (SDEA) já incorporam erros estocásticos e intervalos de confiança, gerando escores em que, com muita frequência, ocorre que nenhuma DMU recebe um escore igual a 100%.

É importante ressaltar que, no contexto deste capítulo, a eficiência é sempre relativa, pois é calculada na amostra. Ao contrário dos exemplos da física e da engenharia, em economia, nas ciências sociais, e nas ciências da saúde, nem sempre é possível saber qual a quantidade máxima que uma DMU pode produzir a partir de um vetor fixo de insumos. Por exemplo: é difícil responder quantas cirurgias, consultas e exames por ano um grande hospital pode produzir. Então, o método recorre aos valores efetivamente observados na amostra, buscando referência virtuosas (*benchmarks*) para inferir o que cada DMU poderia fazer, supondo que ela pudesse mimetizar as melhores práticas efetivamente observadas.

Entre os principais resultados gerados pela DEA, em seus modelos mais básicos, destacamos:

1. escores: usualmente, valores de eficiência técnica, que estão no intervalo entre zero e a unidade ou entre 0% e 100%;

2 *targets* ou alvos: valores ótimos de insumos e de produtos;

2. *peers* ou *benchmarks*: unidades eficientes de referência para as unidades ineficientes;

3. eficiência de escala e eficiência alocativa;

4. contribuição de cada insumo e produto para a eficiência;

5. variação da eficiência e da tecnologia ao longo do tempo (usando MI).

A DEA tem grande e crescente utilização em avaliação de eficiência no Brasil e no exterior, tanto em trabalhos acadêmicos, alguns aqui já citados, como no setor público. Podemos

mencionar, pelo menos, a seguintes instituições públicas brasileiras que fizeram em algum momento, ou ainda fazem, uso dessa metodologia: o Conselho Nacional de Justiça (CNJ) na avaliação da produtividade dos Tribunais Federais, Estaduais e do Trabalho; a Agência Nacional de Telecomunicações (Anatel) no cálculo de produtividade de empresas para modelo de reajuste de preços de serviços de telefonia fixa; a Agência Nacional de Energia Elétrica (Aneel) na avaliação de eficiência de empresas distribuidoras de energia elétrica; a Companhia Energética de Minas Gerais (Cemig) para avaliar a eficiência de custos de empresas; a Secretaria de Estado de Fazenda do Rio de Janeiro (Sefaz-RJ) para avaliar a eficiência de arrecadação; e o Tribunal de Contas da União (TCU) em suas auditorias operacionais.

Análise de Fronteira Estocástica (SFA)

A SFA é um método paramétrico que assume uma relação estocástica entre os insumos e os produtos. Ela assume, *a priori*, a existência de uma função de produção específica subjacente (exemplos: Cobb-Douglas, Translogarítmica) e um processo de geração e dados (DGP) conhecidos, com exceção de um conjunto finito de parâmetros que devem ser estimados a partir dos dados. A SFA é um modelo de regressão obtido por métodos econométricos. Esse modelo foi desenvolvido originalmente, de modo simultâneo, nos trabalhos de Aigner, Lovell e Schmidt (1977)[11], Battese e Corra (1977)[12] e Meeusen e Van den Broeck (1977)[13].

Na forma da função de produção, o modelo mais simples será:

$Y=f(X, \beta)+v-u$.

Em que Y é um vetor de um único produto, ou resultado, observado em cada DMU e X é uma matriz dos vários insumos observados em cada DMU. Então, em um modelo de SFA de função de produção somente, podemos, nos modelos básicos, avaliar DMU produtoras de um único produto a partir de um ou mais insumos. Essa é uma limitação da SFA em relação à DEA que sempre admite múltiplos insumos e múltiplos produtos. Conforme veremos, os modelos de SFA de custos contornam esse problema, se os dados de custos existirem, obviamente. Os modelos de SFA com funções distância, que são de difícil compreensão, e pouco usados, também contornam esse problema[1].

Em termos econométricos[X], a regressão a ser estimada, usualmente por métodos de máxima verossimilhança, será: $Y=X\beta+v-u$. O parâmetro v é o termo usual a aleatório da regressão OLS com distribuição normal e sinal livre. Então, v pode aumentar ou diminuir o produto. O parâmetro $u\geq 0$ (não negativo) mede a ineficiência

X Para uma visão geral sobre econometria, consultar Maddala[6].

produtiva. Como o parâmetro u é sempre não negativo, ele diminuirá o produto se for positivo, e não afetará o produto se for nulo. Caso u seja nulo em uma DMU, a produção será a máxima possível, e a DMU será dita eficiente. Então, uma DMU será eficiente se, e somente se, u=0. Aqui valem as ressalvas e limitações sobre os termos 'eficiente' e 'ineficiente' que fizemos em DEA. A critério do avaliador, ao parâmetro u, é atribuída, usualmente, uma distribuição de probabilidades contínua que assume apenas valores não negativos, como a distribuição seminormal, normal truncada, exponencial ou gama. Em termos teóricos, encontrar u=0 (nulo) é impossível, dado que esse parâmetro tem uma distribuição de probabilidades contínua, e a probabilidade de realização de qualquer ponto isolado é nula. Na prática, isso pode acontecer por causa de aproximações nos cálculos.

Os parâmetros u e v são independentes. Como v pode ser positivo e com módulo maior do que u, é possível que algumas DMU fiquem posicionadas acima da fronteira, que é uma espécie de média das observações, em modelos de OLS.

O modelo estima os valores dos parâmetros β; v; e u. Então, além do valor dos escores de eficiência, os referidos parâmetros devem ser avaliados com os cuidados usuais em modelos de regressão (significância estatística, sinais esperados, magnitude etc.). Dependendo da amostra, pode ser que a regressão não forneça bons resultados, que os parâmetros estimados não sejam adequados e que a utilização de um dado modelo de SFA tenha de ser descartada ou modificada para uso em uma amostra, como pode acontecer em qualquer modelo econométrico.

Na forma de função custo, a formulação será:

$$C=C(Y, w)+v+u.$$

Nesse caso, C são os custos e w é o vetor de custos por unidade (preços de compra) de cada um dos fatores de produção ou insumos. A variável Y agora é uma matriz com todos os produtos de cada DMU, e o modelo acomoda produção múltipla. Valem, para a fronteira de custos, as demais definições e restrições da fronteira de produção. Como o parâmetro u aqui também é não negativo em cada DMU, quanto mais positivo ele for, maior será o custo da DMU e maior a sua ineficiência. Uma DMU será eficiente se, e somente se, u=0.

Os modelos mais básicos de SFA geram os seguintes resultados:

1. escores de eficiência de cada DMU;

2. medidas dos impactos de cada variável dependente (regressores[XI]) na variável de desfecho (elasticidades[14] em modelos com logaritmos das variáveis[XII]);

3. significância estatística dos parâmetros;

4. medidas de retornos de escala;

5. variação da eficiência e da tecnologia ao longo do tempo (em modelos com painel de dados).

Então, como vimos, a SFA é um modelo econométrico especialmente dedicado à mensuração de eficiência econômica. Nesse método, o termo de erro da regressão é dividido entre um componente puramente aleatório e um componente que mede a ineficiência de cada DMU. Como não há garantia de que as regressões apresentem os resultados esperados (significância estatística e sinais corretos dos parâmetros) e demandem treinamento prévio em econometria, a sua aplicação em saúde, no Brasil, tem sido bem mais limitada do que o observado na DEA.

Comparação entre DEA e SFA

É comum, entre iniciantes em avaliação de eficiência, a dúvida sobre qual modelo escolher entre DEA e SFA. Como os modelos são complementares (embora em certo grau também sejam substitutos), aconselhamos, idealmente, o uso concomitante de ambas as metodologias. Caso isso não seja possível, a escolha pode ser auxiliada pela comparação abaixo, que está limitada aos modelos mais simples de ambos os métodos. Existem aprimoramentos consideráveis e uma quase infinidade de opções na literatura.

1. Ao contrário do que ocorre na SFA, a DEA não necessita assumir uma forma funcional (fórmula explícita: Cobb-Douglas, Translogarítmica etc.) de uma função de produção, que transforme insumos em produtos. Por isso, a DEA tem uma grande flexibilidade em relação à SFA.

2. Os modelos básicos de DEA assumem que qualquer desvio em relação à fronteira de produção significa ineficiência. Isso não acontece com a SFA, que tem um parâmetro específico para acomodar erros aleatórios (o parâmetro v). Então, a DEA

XI Os parâmetros dos regressores, usualmente representados pela letra grega β, são equivalentes à declividade, ou coeficiente angular de uma reta ou, mais genericamente, à declividade de uma função afim (***affine function***).

XII Consultar Varian[14] sobre o conceito de elasticidade e outros conceitos microeconômicos.

é mais vulnerável aos erros de medida e aos eventos fortuitos e ruídos nos dados do que a SFA, que distingue ineficiência de erros aleatórios. Os modelos de SDEA mitigaram esse problema[3].

3. A DEA não acomoda *missings* (dados faltantes) e, usualmente, não trabalha de modo simples com painéis desbalanceados (anos faltantes, para algumas DMU, em dados em painel). Alguns modelos e programas de SFA são capazes de contornar esses problemas. Por outro lado, a DEA pode trabalhar em modelos avançados, mas já facilmente disponíveis, com produtos indesejáveis (por exemplo: mortes ou infecções em hospitais), o que não seria possível em modelos comuns de SFA.

   A DEA é muito menos sensível ao tamanho da amostra do que a SFA. Como a SFA é um modelo de regressão, não é possível trabalhar com amostras muito pequenas (menor do que algo em torno de trinta observações). Na DEA, a limitação, para que haja uma boa discriminação (dispersão) dos escores das DMU, advém da relação entre as quantidades de insumos e produtos e a quantidade de DMU. Costuma-se observar uma regra de bolso em que a quantidade de DMU deve ser, ao menos, três vezes maior do que as quantidades somadas de insumos e produtos. Então, na DEA, para um modelo de um insumo e um produto, necessitaríamos, ao menos, de seis DMU, para que não ocorra o problema de que todas as DMU sejam avaliadas como eficientes. Isso ocorre na DEA porque, se existirem muitos insumos e produtos, e forem relativamente poucas as DMU, haverá uma quantidade grande de 'critérios' de avaliação (cada insumo e cada produto seria um critério), e as DMU sempre poderão se especializar em algum insumo ou produto e ter um bom escore, mesmo que tenha mau desempenho nos restantes. Nesse caso, diz-se que ocorre eficiência por virtude de especialização (*by virtue of specialization*). Por exemplo: se quisermos avaliar um conjunto pequeno de alunos de medicina com base em uma quantidade excessivamente grande de matérias, haverá a possibilidade de que cada aluno tenha bom desempenho em algumas matérias, e não tão bom desempenho em outras, de modo muito diferente entre eles, complicando a hierarquização. O modelo chamado de super DEA[15] contorna essa dificuldade, à custa da perda, ou dificuldades de obtenção, de algumas informações relevantes e mesmo da impossibilidade de cálculo dos escores, principalmente em modelos com retornos variáveis de escala.

4. A SFA, por ser um modelo econométrico, permite naturalmente construir testes de hipótese e intervalos de confiança. Os modelos básicos de DEA necessitam de modificações, ou de sofisticações consideráveis, para permitir tais recursos estatísticos encontrados nos modelos SDEA. Por outro lado, a SFA, ao contrário da DEA, é sensível

aos usuais problemas de regressão como multicolinearidade, heteroscedasticidade etc.[6]. Entretanto, por trabalhar com a envoltória (*boundary*) dos dados, os modelos básicos de DEA costumam ser bem mais sensíveis à presença de dados aberrantes (*outliers*) do que a SFA, que trabalha com parâmetros 'médios' (as declividades - os βs) na amostra.

5. A DEA pode ser mais útil que a SFA, se você estiver interessado em um método que permita comparações entre DMU mais próximas umas das outras, para fazer comparações 'locais', ou em comportamentos muito distintos do comportamento médio das DMU. A DEA também extrai mais facilmente as informações das DMU individuais, como valores ótimos de produção e consumo, e referências virtuosas (*benchmarks*), além de permitir visualizar combinações lineares e convexas (médias ponderadas) de DMU. Contudo, se você estiver interessado em comportamentos médios na amostra, pense em SFA.

Em resumo, se o interesse for em alguns detalhes de poucas DMU, usar a DEA pode ser uma opção prática. Caso o interesse seja sobre médias de amostras grandes, a SFA pode ser o modelo de escolha. Se a amostra tiver muito menos de 30 DMU, observadas em um período curto, não há como usar SFA, que é uma regressão, e usar a DEA seria praticamente mandatório. Na presença de múltiplos produtos, e na ausência de dados sobre custos, que impeçam usar uma SFA de custos, a DEA não deve ser descartada. Na presença de dados com qualidade muito duvidosa, ou de *outliers* importantes, que afetam demais os escores das demais DMU, e que não podem ser retirados da amostra sob pena de perda de representatividade, a utilização de DEA, muito sensível a erros de medidas e aos *outliers*, não é aconselhada. Nesse caso, a SFA deveria ser considerada.

Observadas as boas propriedades dos indicadores de eficiência, e escolhidos os modelos de análise, apresentamos um roteiro mínimo para a tarefa de avaliação. Obviamente, trata-se apenas de uma tentativa muito simples para organizar a reflexão de avaliadores e interessados, e não de uma fórmula mágica, ou panaceia, para ser seguida em qualquer situação:

1. Delimite bem o seu problema. Pense sobre os objetivos da avaliação de eficiência, conforme discutido em Marinho e Façanha[16]. Esses objetivos devem convergir com os objetivos organizacionais explícitos ou implícitos. Vale observar que, nem sempre, na prática, objetivos são revelados e que isso não é, necessariamente, um defeito organizacional. Até porque nem sempre é fácil especificar claramente os objetivos de ações, sistemas e programas governamentais. No setor público, particularmente,

é comum que os objetivos não sejam facilmente explicitáveis devido a vários fatores: as diferentes naturezas dos executores das ações e programas governamentais; a multiplicidade de objetivos; a perenidade ou cronicidade dos objetivos e dos problemas, inclusive com a duração das atividades, além de anos fiscais ou mandatos governamentais; as eventuais necessárias transversalidade e multidisciplinaridade dos tratamentos para os problemas enfrentados; as dificuldades de quantificação; e os problemas de interações e superposições de responsabilidades entre programas e ações de governo.

2. Registre os valores observados dos resultados – que podem ser produtos intermediários (*outputs*); ou produtos finais (*outcomes*) – e os insumos e/ou os custos, para cada DMU. Seja cuidadoso, pois essa escolha define os resultados da sua análise em grande extensão, talvez mais do que os métodos de cálculo.

3. Especifique, ou suponha, alguma forma de relação entre insumos e resultados. Essa relação pode ser a tecnologia; ou a correlação (ou a monotonicidade); ou a função de produção; ou a função custo.

4. Faça uma previsão do comportamento eficiente, que permita decidir entre as diversas metodologias disponíveis de avaliação de eficiência (por exemplo: DEA, SFA ou FDH).

5. Para cada DMU, além dos escores, se for possível, calcule as diferenças entre os dados observados e os valores ótimos de insumos e resultados. Os valores ótimos podem ser previstos, na SFA, por uma função de produção, por uma função custo ou pela DEA. Essas diferenças são as ineficiências. É interessante trabalhar com os valores agregados das ineficiências em cada insumo e produto. Esses agregados dão uma ideia do produto potencial e da economia potencial de insumos ou de custos que seriam obtidos se todas as DMU fossem eficientes.

6. Sempre que houver possibilidade, trabalhe com *outcomes* em vez de *outputs*, ou, pelo menos, combine os dois tipos de indicadores. Como assinalamos no capítulo 2, os *outcomes* são os indicadores realmente finalísticos em saúde, que representam uma medida do benefício adicionado pelo tratamento (em sentido amplo) à saúde dos pacientes ou indivíduos, sob a égide de um sistema de saúde. Como exemplo de *outcomes*, temos, entre outros: sobrevida atuarial; Anos de Vida Ajustados por qualidade (*Quality Adjusted Life Years* – QALYS); taxas de mortalidade e de infecção hospitalar (que são *outcomes* indesejados); taxas de sobrevivência (que usaremos no capítulo 7). Os *outputs* são medidas de ati-

vidade, que podem não retratar benefícios para as pessoas. Os exemplos mais comuns são: internações, consultas e exames. Pessoas podem ser reinternadas por terem recebido tratamento inadequado, ou internadas e receberem alta em más condições de saúde; e consultas e exames são procedimentos que podem ser realizados de modo excessivo, sem implicar melhoria da saúde de quem a eles foi submetido. É comum na literatura, até por facilidade de obtenção de dados, entre outras razões, que tais *outputs* sejam utilizados, principalmente em análise de eficiência de hospitais[17], o que, na ausência de *outcomes*, é um fator limitante às análises de eficiência.

Este capítulo fez uma breve incursão aos indicadores, métodos e modelos mais conhecidos e frequentes de avaliação de eficiência econômica. São muitas as opções e as dificuldades para a avaliação de eficiência econômica em saúde. Fundamentalmente, para fazer boas escolhas, é preciso que o avaliador tenha clareza dos objetivos da avaliação, da natureza das instituições que serão avaliadas e do grau de conhecimento que dispõe sobre os diferentes métodos. Indicadores e métodos mal escolhidos, ou mal aplicados, ou mal apresentados, podem esconder mais do que revelar. A busca de atalhos, que podem gerar resultados rapidamente ou de fácil exposição, pode levar a erros crassos com consequências graves sobre políticas, instituições, pessoas e vidas. As coisas certas têm de ser feitas do modo certo.

Referências

1. Bogetoft P, Otto L. Benchmarking with DEA, SFA and R. New York: Springer; 2011.

2. Simar L, Wilson PW. Estimation and inference in two-stage, semiparametric models of production processes. J Econom. 2007;136(1):31-64.

3. Simar L, Wilson PW. Sensitivity analysis of efficiency scores: how to bootstrap in nonparametric frontier models. Manag Sci. 1998;44(1):49-61.

4. Siqueira MM, Araujo CA, Roza BA, Schirmer J. Indicadores de eficiência no processo de doação e transplante de órgãos: revisão sistemática da literatura. Rev Panam Salud Publica. 2016;40(2):90-7.

5. Marinho A. Avaliação da eficiência técnica nos serviços de saúde nos municípios do Rio de Janeiro. Rev Bras Econ. 2003;57(3):515-34.

6. Maddala GS. Introduction to Econometrics. 3rd ed. Great Britain: John Wiley & Sons; 2001.

7. Hillier FS, Lieberman GJ. Introduction to Operations Research. 6th ed. Singapore: McGraw-Hill Inc; 1995.

8. Simar L, Wilson PW. Hypothesis testing in nonparametric models of production using multiple sample splits. J Product Anal. 2020;53(3):287-303.

9. Farrell MJ. The measurement of productive efficiency. J R Stat Soc. Series A (General).1957;120(3):253-90.

10. Charnes A, Cooper WW, Rhodes E. Measuring the Efficiency of Decision Making Units. Eur J Oper Res. 1978;2(6):429-44.

11. Aigner D, Lovell CAK, Schmidt PS. Formulation and estimation of stochastic frontier models. J Econom. 1977;6:21-37.

12. Battese GE, Corra GS. Estimation of a production frontier model: With application to the Pastoral Zone of Eastern Australia. Aust J Agric Econ. 1977;21(3):169-79.

13. Meeusen W, van den Broeck J. Efficiency estimation from Cobb-Douglas production functions with composed error. Int Econ Rev. 1977;18(2):435-44.

14. Varian HR. Microeconomic Analysis. 3rd ed. New York: W.W. Norton & Company; 1992.

15. Andersen e Petersen, 1993.

16. Marinho A, Façanha LOF. Programas sociais: efetividade, eficiência e eficácia, como dimensões operacionais da avaliação. In: Paula LF, Ferreira LR, Assis M, organizadores. Perspectivas para a Economia Brasileira: inserção internacional e políticas públicas. Rio de Janeiro: EdUERJ; 2006. p. 353-368.

17. Almeida Botega L, Andrade MV, Guedes GR. Brazilian hospitals' performance: an assessment of the unified health system (SUS). Health Care Manag Sci. 2020;23(3):443-52.