EFEITO MÉDIO DO TRATAMENTO (ATE): MÉTODOS PARA ESTIMAÇÃO E APLICAÇÕES EM ANÁLISE DE CAUSALIDADE

Gabriel Leite Mariante

**Efeito Médio do Tratamento (ATE): Métodos para
Estimação e Aplicações em Análise de Causalidade**

Aluno(s): Gabriel Leite Mariante

Orientador(es): Cristiano Augusto Coelho Fernandes

Trabalho apresentado com requisito parcial à conclusão do curso de Engenharia Elétrica na Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Brasil.

Agradecimentos

Como este trabalho marca o fim de um ciclo importante em minha vida, creio ser importante agradecer não só àqueles que contribuíram diretamente pra sua elaboração, mas também a todos que, de alguma forma, foram importantes para mim durante este período marcante.

Primeiro, agradeço a meus pais, Mariana e Armando, pela educação, carinho, apoio e incentivo infinitos. A eles, devo todas as oportunidades que tive na vida.

Agradeço ao Professor Cristiano Fernandes, meu orientador, pelo cuidadoso aconselhamento durante a realização deste trabalho e por todo o direcionamento acadêmico, muito importante na reta final do meu curso de Graduação.

Agradeço a todos os professores que tive na vida, em especial os do Colégio Santo Inácio, da PUC-Rio e da University of California, Santa Barbara. São eles os responsáveis mais diretos por tudo aquilo que aprendi ao longo de todos os anos de estudo.

Agradeço à PUC-Rio e a todos os seus funcionários pelo excelente ambiente acadêmico, tão propício ao estudo e ao aprendizado.

Aos meus queridos amigos dos tempos de colégio e àqueles igualmente queridos com os quais a PUC-Rio me presenteou durante estes cinco anos e meio, agradeço pela amizade, companheirismo e convivência estimulante. Com eles, compartilhei os bons (e também os maus!) momentos destes anos universitários. Agradeço em especial ao Bernardo que, além de fazer parte deste grupo, contribuiu de maneira relevante para a elaboração deste trabalho.

Este trabalho tem como objetivo a análise teórica de diversas metodologias estatísticas e econométricas de estimação do Efeito Médio do Tratamento (ATE) e sua aplicação em análise de causalidade através de um estudo de caso com dados de um projeto de implementação de uma política educacional.

Em um experimento perfeitamente controlado e aleatorizado, há homogeneidade entre os grupos tratado e não tratado, podendo o ATE ser estimado pela simples diferença entre as médias dos dois grupos. No entanto, na prática, tais experimentos são raros e é conveniente que haja metodologias estatísticas para relaxar a hipótese de aleatoriedade perfeita. Se assumirmos que, apesar da não existência de aleatoriedade na atribuição do tratamento, é possível explicar tal atribuição a partir de um conjunto de variáveis observadas, pode-se estimar o ATE de maneira consistente através de uma regressão múltipla da variável resultado na variável tratamento e nas demais variáveis observadas. Sob a mesma hipótese, pode-se também estimar o ATE a partir da estimativa do escore de propensão, um método de pareamento de indivíduos baseado na estimativa da probabilidade individual de seleção.

Finalmente, se não houver aleatoriedade e a atribuição do tratamento não for explicada, pode-se obter uma estimativa consistente do ATE a partir do método de Variáveis Instrumentais. Para isto, é necessário apenas que alguma das variáveis observadas na população, chamada de instrumento, possua duas condições específicas: seja correlacionada com o tratamento e exógena ao modelo de regressão original.

As metodologias apresentadas são aplicadas aos dados de uma política educacional implementada nos Estados Unidos nos anos 80 que buscava quantificar o efeito da redução do número de alunos por turma no aprendizado. Os resultados mostram que há um efeito benéfico estatisticamente significativo, porém, pequeno, o que leva a um questionamento da relevância e da viabilidade prática desta política.

Palavras-chave: Efeito Médio do Tratamento (EMT); causalidade; regressão, escore de propensão; Projeto Tennessee STAR

Average Treatment Effect (ATE): Estimation Methods applied to Causality Analysis.

Abstract

This work analyses and discusses several statistical and econometric methodologies for estimation of the Average Treatment Effect (ATE) and its application to causality analysis through a case study with data from an educational policy.

In a perfectly controlled and randomized experiment, there is homogeneity of individuals in the treated and non-treated groups. Therefore, the ATE can be estimated simply by the difference between the sample means of the two groups. However, such experiments do not happen very often and it is convenient to have statistical methodologies to relax the hypothesis of perfect randomness. If we assume that, despite the non-randomness of treatment, it is possible to explain its selection through a set of observed variables, we can consistently estimate the ATE through a linear regression of the result variable on the treatment variable and on the other observed explanatory variables. Under the same hypothesis, we can estimate the ATE from the estimation of the propensity score, which is a matching method based on the estimation of the individual probability of treatment selection.

Finally, if there is no randomness and we cannot explain treatment selection, we can obtain a consistent estimation for the ATE through the method of Instrumental Variables. The only condition is that one of the observed variables is both correlated with treatment selection and uncorrelated with the linear regression error term.

The methodologies were applied to data from an educational policy implemented in the United States in the 1980's. Its goal was to quantify the effect of class size reduction in student learning. The results show a statistically significant positive effect. However, such effect is very small, which leads to a debate on the relevance and practical viability of such policy.

Keywords: Average Treatment Effect (ATE); causality; regression; propensity score; Tennessee STAR Project

1. Introdução	1
1.1. Efeito Médio do Tratamento (ATE)	1
1.2. Considerações sobre a notação utilizada	1
1.3. Organização dos capítulos	2
2. Metodologias de Estimação do ATE	3
2.1. Problemas de inferência causal de definição do ATE	3
2.1.1. Contrafactuais e o Problema Fundamental da Inferência Causal	3
2.1.2. Viés de seleção	4
2.2. Estimação através de regressão direta	6
2.3. Estimação através do Escore de Propensão	9
2.3.1. Estimação direta do ATE	10
2.3.2. Estimação via regressão	11
2.4. Estimação através do método de Variáveis Instrumentais	13
2.4.1. Introdução à regressão com Variáveis Instrumentais	13
2.4.2. Aplicação à estimação do ATE	17
2.5. Resumo das metodologias de estimação apresentadas	19
3. Estudo de Caso	21
3.1. O Projeto Tennessee STAR	21
3.2. Descrição do banco de dados	22
3.2.1. Descrição das variáveis	23
3.2.2. Estatísticas descritivas	23
4. Resultados	25
4.1. Método 1	25
4.2. Método 2	25
4.3. Método 3	26
4.4. Método 4	27
4.5. Método 5	27
4.6. Método 6	28
4.7. Método 7	28
4.8. Resumo dos resultados	30
5. Conclusão	32

6. Referências Bibliográficas	34
Anexo	35

1. Introdução

O conceito de causalidade está presente nas mais diversas áreas do conhecimento e, mais especificamente, em diversas modalidades de análise estatística e econométrica. Trata-se de um conceito amplo e que, segundo Angrist e Pischke (2009), tem significados diferentes para pessoas diferentes. Apesar disso, pode-se dizer que, de uma maneira geral, um dado evento é a causa de um dado resultado se sua ocorrência foi determinante para a existência deste, isto é, sem o dado evento, o resultado esperado teria sido diferente do encontrado.

Medir e quantificar relações de causalidade é fundamental e, muitas vezes, não trivial. Luiz e Struchiner (2002) exemplificam a importância de medidas estatísticas e econométricas de efeitos causais na área de Epidemiologia. Nesta área, a análise causal é utilizada para responder questões acerca da eficácia de determinados medicamentos ou das relações entre uma determinada doença e seus possíveis agentes causadores.

Heckman (2005) aborda uma aplicação mais próxima da estudada neste artigo: a utilização de medidas matemáticas de causalidade para avaliar e planejar políticas públicas e intervenções governamentais na sociedade. No campo da avaliação, é importante buscar metodologias que permitam afirmar que um dado resultado foi de fato causado por uma dada política pública e não por alguma outra mudança social não relacionada à intervenção. Já no campo do planejamento, deve-se enfrentar o problema de prever os resultados de uma política ainda não implementada. Muitas vezes, a aplicação de uma nova política é baseada na existência de políticas semelhantes e bem-sucedidas em outras regiões. No entanto, como afirmar que os resultados serão os mesmo se condições sociais relevantes para o sucesso da intervenção podem ser diferentes? Diversas metodologias estatísticas e econométricas podem ser utilizadas para responder a estas questões fundamentais. Algumas delas serão expostas nas próximas seções.

1.1. Efeito Médio do Tratamento (ATE)

Neste trabalho abordamos de maneira pormenorizada uma das principais medidas da relação causa-efeito na literatura estatística e econométrica, o chamado Efeito Médio do Tratamento, mais comumente conhecido por sua sigla em inglês ATE (Average Treatment Effect).

No contexto estudado, a palavra “tratamento” é usada para definir genericamente qualquer tipo de ação à qual uma parte de uma população foi submetida e outra não. A realização de uma dada política pública visando um determinado grupo de pessoas na sociedade, a administração de um remédio para parte dos pacientes em um estudo clínico ou algum hábito individual particular a respeito do qual se deseja estudar possíveis consequências (como fumar, por exemplo) são exemplos do que podemos chamar de tratamento.

O ATE é obtido em geral a partir de diferentes técnicas econométricas e estatísticas como, por exemplo, modelos de regressão linear ou não linear, pareamento de unidades e variáveis instrumentais. Seu objetivo é quantificar o resultado médio da exposição de indivíduos de uma dada população a um dado tratamento. Dito de outra maneira, para um dado tratamento que, supõe-se, é responsável por alguma consequência quantificável para os indivíduos de uma dada população, o ATE quantifica a diferença média entre ser ou não ser submetido ao tratamento em questão.

1.2. Considerações sobre a notação utilizada

Este trabalho utiliza diversas equações que representam modelos de regressão linear do tipo:

$$y = \beta_0 + \rho w + \beta x + u$$

A seguinte convenção será adotada:

- Variáveis, que podem assumir valores distintos para os indivíduos da população, serão representadas por letras latinas em fonte diferente e itálico, por exemplo: w
- Coeficientes de um modelo de regressão linear são representados por letras gregas em fonte diferente e em itálico, por exemplo: β_0

- Vetores são representados em fonte diferente, itálico e negrito, utilizando letras latinas para vetores que representam conjuntos de variáveis e gregas para vetores que representam coeficientes de um modelo de regressão, por exemplo: β e x
- Estimadores de uma variável ou de um coeficiente são representados pelo acréscimo de um acento circunflexo sobre a dada variável ou coeficiente, por exemplo: $\hat{y}$ e $\hat{\beta}_0$
- A média amostral de uma variável é representada através do acréscimo de uma barra vertical sobre a mesma, ou seja: $\bar{x} = \sum_{i=1}^n \frac{x_i}{n}$
- Não é especificado, por falta de necessidade das contas aqui realizadas, se os vetores são representados como linhas ou colunas. É suficiente definir que a multiplicação de um vetor por outro aqui é sempre representada como um produto interno, a saber: $ab = \sum_{i=1}^n a_i b_i$.

1.3. Organização dos capítulos

O capítulo 2 apresenta o desenvolvimento teórico deste trabalho. Inicialmente, são apresentados alguns problemas básicos na estimação do ATE e, a seguir, são expostas três metodologias para sua estimação, cada qual com algumas variações, perfazendo um total de sete métodos de estimação. Todas elas são acompanhadas de considerações teóricas a respeito da ideia geral, fórmulas matemáticas para sua aplicação e hipóteses que devem ser respeitadas para sua validade.

O capítulo 3 apresenta os dados utilizados em um estudo de caso no qual as metodologias desenvolvidas no capítulo 2 serão aplicadas. Trata-se do projeto Tennessee STAR, uma política educacional implantada durante os anos 1980 em escolas públicas no estado americano do Tennessee. O objetivo da política foi verificar se a diminuição do tamanho das turmas nos anos iniciais do ensino primário tinha impacto significativo no aprendizado dos alunos.

O capítulo 4 apresenta os resultados da aplicação das metodologias desenvolvidas neste trabalho aos dados referentes ao projeto Tennessee STAR.

O capítulo 5 conclui e o capítulo 6 enumera referências bibliográficas citadas neste trabalho e recomendadas ao leitor para uma consulta mais aprofundada sobre os temas aqui abordados.

2. Metodologias de Estimação do ATE

Neste capítulo serão apresentados os principais problemas relacionados à estimação do ATE e algumas metodologias para efetivá-lo.

Por causa do chamado Problema Fundamental da Inferência Causal, apresentado na seção 2.1.1, muitas das metodologias utilizadas para estimar o ATE são construídas a partir de hipóteses acerca dos dados analisados. Algumas das hipóteses são razoáveis e, em geral, se verificam com facilidade em estudos com dados reais, enquanto outras são muito restritivas e dificilmente são verificadas.

Deve-se ressaltar que a verificação prévia de que os dados analisados estão de acordo com as hipóteses propostas é fundamental para a correta estimação do ATE. A violação das hipóteses estabelecidas invalida os cálculos estatísticos e o resultado final da estimação. O leitor verá que existem diferentes relações de custo-benefício entre a complexidade teórica da metodologia e o quão restritivas são as hipóteses necessárias para aplicá-la.

É importante ressaltar que este trabalho não é exaustivo. Há uma enorme diversidade de métodos para estimação do ATE e uma extensa literatura sobre o assunto. As referências no capítulo 6 devem ser consultadas caso o leitor deseje se aprofundar no tópico.

2.1. Problemas de inferência causal e definição do ATE

Para ilustrar as demonstrações subsequentes, suponha uma população de N indivíduos onde se quer estimar o valor do ATE para um dado tratamento. Naturalmente, uma parte destes indivíduos foi submetida ao tratamento em questão e outra não o foi.

Podemos então definir uma **variável tratamento** que, associada a cada indivíduo da população, define se o indivíduo foi ou não tratado. Seja w uma variável binária, tal que cada indivíduo tenha um valor associado de w . Para todos os indivíduos que foram submetidos ao tratamento, $w = 1$. Para os indivíduos que não foram submetidos ao tratamento, $w = 0$.

Além disso, define-se uma **variável resposta** que, também associada a cada indivíduo da população, supõe-se que tenha uma relação causal com o tratamento w e que se deseja quantificar esta relação. A variável resposta será chamada de y .

Como parte dos indivíduos foi submetida ao tratamento e outra não, pode-se então definir, a partir da variável resposta, o par (y_0, y_1) , onde y_0 denota a variável resposta para casos sem tratamento (ou seja, com $w = 0$) e y_1 denota a variável resposta para casos com tratamento ($w = 1$).

A população de N indivíduos, portanto, é descrita por (y_0, y_1, w) . Podemos agora definir o efeito médio do tratamento (ATE) como a diferença média entre ser ou não ser tratado, ou seja, como o valor esperado da diferença entre y_1 e y_0 .

$$ATE = E[y_1 - y_0] \quad (1)$$

2.1.1. Contrafactuais e o Problema Fundamental da Inferência Causal

Há uma forte limitação de ordem observacional para o cálculo do ATE chamada por Holland (1986) de “Problema Fundamental da Inferência Causal”. Como definido, para um dado indivíduo, a variável w , que define se este indivíduo foi ou não submetido ao tratamento, assume um único valor igual a 0 (caso o indivíduo não tenha sido tratado) ou igual a 1 (caso ele tenha sido tratado). Desta maneira, para cada uma das unidades observadas, apenas será possível observar um dos dois valores do par (y_0, y_1) . A observação de ambos é, naturalmente, impossível. Os indivíduos tratados terão seus valores

para y_1 observados, mas nada sabemos sobre seus valores para y_0 , e o contrário ocorrerá com os indivíduos não-tratados.

Como não se pode observar simultaneamente o par (y_0, y_1) para um mesmo indivíduo, não é possível calcular a diferença individual entre ser ou não ser tratado, pois falta um dos valores. Será visto mais a frente que, em experimentos aleatórios, ou seja, aqueles em que a atribuição ou não do tratamento é decidida aleatoriamente, o Problema Fundamental é facilmente contornado. A aleatorização implica em uma homogeneidade de indivíduos nos dois grupos (o tratado e o não tratado), ou seja, em média, todas as demais características estarão igualmente distribuídas nos dois grupos. Em situações em que não há aleatoriedade de tratamento, porém, faz-se necessária a utilização de técnicas estatísticas e econométricas para contornar esta limitação observacional.

Muitas das técnicas para contornar este problema envolvem a estimação individual, para cada unidade, do valor que não foi observado. Isto pode ser feito através de modelos de regressão utilizando outras variáveis características dos indivíduos observados (como faremos na seção 2.2 e 2.4) ou através de métodos de pareamento de unidades (como faremos na seção 2.3). A este valor não observado (isto é, y_0 para os indivíduos tratados e y_1 para os indivíduos não tratados) dá-se o nome de **contrafactual**.

Heckman (2005) também se refere aos valores contrafactuais através do termo “hipotéticos”. Como o valor dos contrafactuais não pode ser observado, pode-se imaginá-los como sendo o valor observado da variável resultado em um cenário hipotético em que os indivíduos não tratados teriam sido tratados e vice-versa.

2.1.2. Viés de seleção

Como vimos na seção anterior, a não observação de um dos valores do par para cada (y_0, y_1) faz com que seja necessária a estimação de valores hipotéticos através de contrafactuais. Agora, veremos que, sob uma hipótese muito restritiva, é possível estimar o ATE através da simples diferença de médias entre os grupos dos indivíduos tratados e não tratados, sem a necessidade de estimar explicitamente valores para os contrafactuais.

Hipótese 1: a variável tratamento w é estatisticamente independente do par de variáveis respostas (y_0, y_1)

Se vale a Hipótese 1, podemos estabelecer relações simples que mostram que a diferença observada nas médias dos grupos é um bom estimador para o ATE. Começemos por escrever a observação da variável resultado y em função de w e do par (y_0, y_1) :

$$y = (1 - w)y_0 + wy_1 = y_0 + w(y_1 - y_0) \quad (2)$$

Ou seja, se $w = 1$ observa-se y_1 e se $w = 0$ observa-se y_0 . Como vale a Hipótese 1, a variável resultado e a variável tratamento são independentes e, então, podemos aplicar o valor esperado condicionado aos dois possíveis valores de w e escrever:

$$E[y|w = 1] = E[y_1|w = 1] = E[y_1] \quad (3)$$

$$E[y|w = 0] = E[y_0|w = 0] = E[y_0] \quad (4)$$

Mas, pela equação (1) que define o ATE, podemos também escrever:

$$ATE = E[y_1 - y_0] = E[y_1] - E[y_0] = E[y_1|w = 1] - E[y_0|w = 0] \quad (5)$$

O lado direito da equação (5) é exatamente a diferença entre as médias dos valores observados no experimento, isto é, os valores de y_1 para os indivíduos tratados ($w = 1$) e de y_0 para os indivíduos não tratados ($w = 0$). Desta maneira, podemos estimar de maneira não-viesada o ATE através da diferença entre a média observada do resultado no grupo dos indivíduos que receberam tratamento e a média

observada do resultado no grupo dos indivíduos que não receberam tratamento, sem a necessidade de estimativa de contrafactuais. Algebricamente:

$$ATE = \bar{y}_1 - \bar{y}_0 \quad (6)$$

No entanto, de um modo geral, a Hipótese 1 só se verifica em experimentos controlados onde a aleatoriedade do tratamento é imposta aos indivíduos. Um exemplo de situação em que ela é respeitada é um teste clínico para medir a eficácia de um dado remédio. Supondo que a seleção dos indivíduos a quem a droga será administrada seja aleatória e controlada, então a estimação do efeito do remetido é feita a partir da diferença entre as médias observadas do grupo tratado e do grupo não tratado.

Este fato é intuitivo, pois pode-se imaginar que, em um experimento aleatorizado, a princípio, a única diferença relevante entre os indivíduos de um grupo e de outro é justamente o fato de terem ou não sido selecionados para o tratamento. Assim, medindo a diferença entre as médias, estamos de fato medindo o efeito do tratamento, e não de outras variáveis ignoradas pelo modelo e que podem influenciar o resultado.

Na maioria dos casos, porém, a Hipótese 1 é violada. Angrist e Pischke (2009) e Luiz e Struchiner (2002) citam o exemplo de um estudo em que se deseja obter a estimação do efeito do fumo na incidência de doenças pulmonares a longo prazo. Neste caso, o tratamento ao qual parte dos indivíduos é submetida é o fumo durante um longo período de tempo. No entanto, por razões éticas, é inviável a realização de um experimento controlado em que indivíduos são aleatoriamente selecionados para fumar durante um período longo de tempo. Analogamente, como exemplifica Heckman (2005), em muitas políticas públicas relacionadas, por exemplo, a programas sociais, a participação ou não de um dado indivíduo é voluntária. Assim sendo, fatores que levam um dado indivíduo a se inscrever em um dado programa são externos e ignorados pelo modelo, quebrando a aleatoriedade suposta pela Hipótese 1. Luiz e Struchiner (2002) ainda citam o frequente caso em que a análise científica de efeito do tratamento é feita *a posteriori*, isto é, depois que o resultado já foi alcançado. Nestas situações, a atribuição do tratamento já foi feita antes da análise e não há, portanto, qualquer controle sobre ela. Muitas vezes, não se sabe se o tratamento foi ou não aleatorizado entre os indivíduos da população estudada.

Nos casos em que a seleção dos indivíduos a serem submetidos ao tratamento não é controlada e aleatória, não valem as equações (3), (4) e (5). Consequentemente, a estimativa do ATE através da simples diferença entre as médias dos grupos é viesada. Como este viés é baseado no método de seleção de indivíduos, dá-se a ele o nome de **viés de seleção**.

A ideia por trás do conceito de viés de seleção é a de que o resultado observado do tratamento pode ser consequência não apenas do tratamento em si, mas de outras variáveis ignoradas pelo modelo e que também influenciam na seleção dos indivíduos para serem submetidos ao tratamento.

Angrist e Pischke (2009) exemplificam o efeito do viés de seleção através de um estudo realizado nos Estados Unidos que desejava medir o efeito da hospitalização de indivíduos em seu estado geral de saúde. Os dados coletados no estudo mostraram que um dado indicador de saúde individual geral (que, aqui, é a variável resultado) de pessoas após serem hospitalizadas era, em média, significativamente pior do que o de pessoas que nunca foram internadas em hospitais. Uma simples estimação do ATE por diferença das médias, portanto, levaria a uma conclusão errônea de que se internar em um hospital tem um efeito negativo na saúde de um indivíduo. O viés de seleção fica evidente neste exemplo, pois, obviamente, indivíduos são hospitalizados justamente porque possuem um estado de saúde prévio pior do que indivíduos que não precisam ser internados. A Hipótese 1, portanto, é violada e não há independência entre o tratamento e o resultado.

As seções 2.2, 2.3 e 2.4 abordam maneiras de estimar o ATE em experimentos em que a seleção de indivíduos para tratamento não é aleatória e, portanto, há viés de seleção na diferença das médias.

2.2. Estimação através de regressão direta

Para a aplicação deste método, além das variáveis introduzidas nas seções anteriores, é necessária a existência de um vetor de n variáveis independentes observadas para todos os indivíduos do banco de dados: $[x_1, x_2, \dots, x_n]$ que, a partir de agora, será representado por x . Com este acréscimo, a população agora é descrita por (y_0, y_1, w, x) .

Na seção 2.1.2 argumentamos que, em experimentos não aleatórios, a Hipótese 1 dificilmente se verifica. Para contornar este problema, vamos agora permitir a correlação entre a variável de seleção w e o par referente às respostas ao tratamento (y_0, y_1) . Uma nova hipótese, bem menos restritiva, deverá ser verificada para validar os cálculos subsequentes.

Hipótese 2: $E[y_0|w, x] = E[y_0|x]$ e $E[y_1|w, x] = E[y_1|x]$

A Hipótese 2, que será fundamental para a sequência deste trabalho, pode ser interpretada como a independência entre w e o par (y_0, y_1) dado um vetor de variáveis observadas x .

Como já visto, em casos em que a Hipótese 1 não é respeitada, a seleção dos indivíduos para tratamento não é aleatória. A ideia principal da Hipótese 2 é a de que as variáveis observadas presentes em x contêm informações a respeito dos indivíduos da população que explicam a atribuição do tratamento aos indivíduos. Assim, controlando por um vetor de variáveis explicativas x , existe independência entre o valor esperado do par (y_0, y_1) e w .

Empiricamente, a verificação ou não desta hipótese fundamental depende dos dados aos quais se tem acesso. Se o banco de dados observado contém variáveis com informação o suficiente para que seja possível estabelecer as causas da atribuição do tratamento, isto é, se o conteúdo de x explicar w , a Hipótese 2, em geral, se verifica. Se o banco de dados for muito pequeno ou incompleto, porém, esta hipótese torna-se muito restritiva e não se verifica.

Se vale a Hipótese 2, podemos então escrever:

$$E[y_1|x, w = 1] - E[y_0|x, w = 0] = E[y_1|x] - E[y_0|x] = E[y_1 - y_0|x] = ATE(x) \quad (7)$$

Isto significa que podemos então estimar o ATE em função de x através das diferenças observadas (lado esquerdo da equação) das médias condicionais ao vetor de variáveis independentes x .

Se o número de variáveis presentes em x que explicam de maneira suficiente a atribuição do tratamento for pequeno e o banco de dados tiver um volume de dados muito grande, é possível estimar o ATE de maneira simples através da lei das expectativas iteradas, fazendo a média condicional a todos os possíveis valores assumidos pelas variáveis em x :

$$E[E[y_1 - y_0|x]] = E[y_1 - y_0] = ATE \quad (8)$$

No entanto, no caso geral, é impossível o cálculo de uma média condicional a todos os possíveis valores assumidos pelas variáveis em x . Desta forma, um modelo de regressão deve ser construído.

Começamos esta construção escrevendo cada elemento do par de variáveis resposta (y_0, y_1) como uma soma entre seu valor esperado (constante) e um termo variável com média nula:

$$y_1 = E[y_1] + v_1 \quad (9)$$

$$y_0 = E[y_0] + v_0 \quad (10)$$

$$E[v_1] = E[v_0] = 0 \quad (11)$$

Podemos então reescrever a equação (2) com esta nova notação para as variáveis resposta:

$$y = y_0 + w(y_1 - y_0) = E[y_0] + v_0 + w[E[y_1] + v_1 - (E[y_0] + v_0)] \quad (12)$$

Rearranjando os termos:

$$y = E[y_0] + w(E[y_1] - E[y_0]) + v_0 + w(v_1 - v_0) \quad (13)$$

Aplica-se então o valor esperado condicionado a w e x dos dois lados:

$$\begin{aligned} E[y|w, x] &= E[E[y_0]|w, x] + E[w(E[y_1] - E[y_0])|w, x] + E[v_0|w, x] + E[w(v_1 - v_0)|w, x] \\ &= E[y_0] + wE[y_1 - y_0] + E[v_0|w, x] + wE[v_1 - v_0|w, x] \end{aligned} \quad (14)$$

Neste ponto é conveniente, por simplicidade do raciocínio, assumir a seguinte hipótese, não muito restritiva.

Hipótese 3: $E[v_0|x] = E[v_1|x]$

Já sabemos que, por construção, os termos do par (v_0, v_1) têm média nula. A Hipótese 3 assume que, condicionado aos valores das variáveis em x , este par tem o mesmo valor esperado. Esta hipótese não é muito restritiva e, além disso, será visto mais a frente que ela pode ser relaxada com uma simples alteração na estimativa final.

Se valem as hipóteses 2 e 3, então:

$$E[v_1 - v_0|w, x] = E[v_1 - v_0|x] = 0 \quad (15)$$

E então:

$$E[y|w, x] = E[y_0] + wE[y_1 - y_0] + E[v_0|x] \quad (16)$$

Seja $h_0(x)$ um vetor função das variáveis independentes observadas x . Um exemplo simples sem perda no raciocínio é assumir a função identidade, onde $h_0(x) = x$. Na verdade, assumir esta simplificação implica em assumir que o vetor x já inclui quaisquer transformações feitas nas variáveis originais observadas e comumente empregadas em modelos de regressão linear (por exemplo, tomar o logaritmo de uma variável). Adotando esta simplificação, podemos escrever o valor esperado condicional do termo v_0 como um termo constante e outro dependente de x :

$$E[v_0|x] = \eta_0 + \beta_0 x \quad (17)$$

Podemos então reescrever a equação (16):

$$E[y|w, x] = E[y_0] + \eta_0 + wE[y_1 - y_0] + \beta_0 x \quad (18)$$

Somando os termos constantes, seja:

$$\gamma_0 = E[y_0] + \eta_0 \quad (19)$$

E, renomeando o termo que representa o ATE, seja:

$$E[y_1 - y_0] = \alpha \quad (20)$$

A equação para o valor esperado da variável resposta fica:

$$E[y|w, x] = \gamma_0 + \alpha w + \beta_0 x \quad (21)$$

Portanto, a equação (21) pode ser estimada através de uma regressão simples da variável resposta y nas n variáveis observadas contidas em x mais a variável binária w , que representa o tratamento. O efeito do tratamento será, então, consistentemente estimado pelo coeficiente encontrado nesta regressão múltipla para a variável w .

$$A\hat{T}E = \hat{\alpha} \quad (22)$$

Com o modelo mais simples de regressão já construído, podemos ir um pouco além e relaxar a Hipótese 3. Se não vale a igualdade de valores esperados condicionais a x para os elementos do par (v_0, v_1) , devemos voltar à equação (14):

$$\begin{aligned} E[y|w, x] &= E[y_0] + wE[y_1 - y_0] + E[v_0|w, x] + wE[v_1 - v_0|w, x] \\ &= E[y_0] + wE[y_1 - y_0] + E[v_0|x] + wE[v_1 - v_0|x] \end{aligned} \quad (23)$$

Podemos utilizar o mesmo argumento utilizado para termo v_0 para reescrever o valor esperado condicional do termo v_1 como:

$$E[v_1|x] = \eta_1 + \beta_1 x \quad (24)$$

Então vale:

$$\begin{aligned} E[y|w, x] &= E[y_0] + wE[y_1 - y_0] + \eta_0 + \beta_0 x + w(\eta_1 + \beta_1 x) - w(\eta_0 + \beta_0 x) \\ &= E[y_0] + \eta_0 + w(E[y_1 - y_0] + \eta_1 - \eta_0) + \beta_0 x + wx(\beta_1 - \beta_0) \end{aligned} \quad (25)$$

Seja:

$$\gamma_0 = E[y_0] + \eta_0 \quad (26)$$

$$\alpha = E[y_1 - y_0] + \eta_1 - \eta_0 \quad (27)$$

$$\pi = \beta_1 - \beta_0 \quad (28)$$

Então a equação (25) pode ser escrita como:

$$E[y|w, x] = \gamma_0 + \alpha w + \beta_0 x + \pi x w \quad (29)$$

A diferença entre esta regressão e a regressão encontrada assumindo a validade da Hipótese 3 (equação (21)) é um termo de interação entre a variável binária w e as variáveis observadas em x . Sem perda, subtraímos de todas as variáveis em x o seu valor médio. Não conhecemos as médias populacionais das variáveis x , mas Wooldridge (2010) mostra que os efeitos da utilização da média amostral ao invés da média populacional desconhecida são desprezíveis para o resultado da estimação.

O modelo final de regressão fica então:

$$E[y|w, x] = \gamma_0 + \alpha w + \beta_0 x + \delta(x - \bar{x})w \quad (30)$$

Se não vale a Hipótese 3, então, o modelo que estima o efeito do tratamento deve ser uma regressão da variável resposta em w , nas variáveis observadas x e na interação entre w e x subtraído dos seus valores médios. O efeito médio do tratamento como função de x é, então estimado por:

$$A\hat{T}E(x) = \hat{\alpha} + \hat{\delta}(x - \bar{x}) \quad (31)$$

Uma vantagem deste método é o fato de que pode-se obter a média amostral do efeito do tratamento em qualquer subgrupo de interesse da população como, por exemplo, só naqueles indivíduos que de fato foram tratados, dando origem a outro estimador comum na literatura, o Efeito Médio nos Tratados (Wooldridge (2010) e Heckman (2005) exemplificam aplicações práticas do Efeito Médio nos Tratados). Fazendo a média em toda a população, tem-se, naturalmente:

$$A\hat{T}E = \hat{\alpha} \quad (32)$$

2.3. Estimação através do Escore de Propensão

O conceito de Escore de Propensão (mais comumente encontrado na literatura com sua nomenclatura em inglês: Propensity Score) foi introduzido na literatura sobre análise de causalidade e efeitos de tratamento por Rosenbaum e Rubin (1983).

O Escore de Propensão se enquadra dentro da categoria mais geral de método de “matching”, ou pareamento de indivíduos. A ideia desta categoria de métodos é comparar indivíduos com características semelhantes cuja única diferença relevante é ter ou não sido tratado. Desta maneira, a diferença na variável resposta entre eles se deve apenas ao tratamento, e não a outras variáveis que poderiam servir como confundimento.

Em um método de pareamento de indivíduos em que um foi tratado e outro não, cada um dos indivíduos pode ser interpretado como o contrafactual de seu par. O valor da variável resposta do indivíduo que foi tratado é igual ao valor hipotético da resposta do indivíduo que não foi tratado, caso tivesse sido, e vice-versa. Dito de outra maneira, o valor de y_0 , que é observado apenas para o indivíduo do par que não foi tratado, vale para os dois, o mesmo ocorrendo com y_1 .

Nas palavras de Luiz e Struchiner (2002), o escore de propensão “é definido como a probabilidade condicional de designação de um particular tratamento dado um vetor de co-variáveis observadas”.

Desta maneira, dada a não aleatoriedade da atribuição de tratamento (ou seja, assim como na seção anterior, não vale a Hipótese 1), obter o escore de propensão significa estimar a probabilidade de que um indivíduo seja selecionado para o tratamento a partir do vetor de variáveis observadas x . Para tal, naturalmente, é preciso que as variáveis presentes em x sejam suficientes para explicar a atribuição de tratamento, representada por w . Ou seja, tal como na seção anterior, é necessário também que valha a Hipótese 2.

Estimar a probabilidade de seleção de cada indivíduo a partir das variáveis presentes em x significa estimar a probabilidade de que, para cada indivíduo, $w = 1$. Podemos então definir:

$$p(x) = \Pr(w = 1|x) \quad (33)$$

Uma maneira eficiente de realizar tal estimativa é através de um modelo de regressão com w como variável dependente e o conjunto de variáveis em x como variáveis explicativas. Como, por definição, $0 \leq p(x) \leq 1$, um modelo de regressão linear pode não ser adequado, pois seria possível estimar valores fora do intervalo $[0,1]$. Desta maneira, modelos de regressão do tipo logit ou probit são mais adequados, pois garantem que a estimativa da variável dependente estará entre 0 e 1 (como mostra a Figura 1).

Figura 1 - Distribuição de probabilidade acumulada de dados artificialmente gerados a partir de um modelo logit. Observa-se que os valores do eixo x estão sempre entre 0 e 1, fazendo com que este seja um modelo adequado para a estimação de uma variável binária.

Wooldridge (2010) argumenta que os modelos logit e probit tendem a apresentar resultados quase idênticos na estimação final do ATE. Neste trabalho, optou-se por utilizar um modelo do tipo logit. Assim, definimos a estimativa do escore de propensão como:

$$\hat{p}(x) = \frac{1}{1 + e^{\beta_0 + \beta x}} \quad (34)$$

Veremos a seguir duas maneiras de estimar o ATE a partir do escore de propensão.

2.3.1. Estimação direta do ATE

A primeira forma de estimação do ATE a partir do escore de propensão decorre do fato de que o ATE pode ser escrito como uma função de $p(x)$. Assim, se obtivermos uma estimativa $\hat{p}(x)$ a partir do modelo logit apresentado anteriormente, podemos estimar diretamente o valor do ATE.

De fato, será mostrado que:

$$ATE = E \left[\frac{(w - p(x))y}{p(x)(1 - p(x))} \right] \quad (35)$$

Para tal, partiremos da equação (2):

$$y = (1 - w)y_0 + wy_1 \quad (36)$$

Substituindo o valor de y no numerador da equação (35), obtém-se:

$$\begin{aligned} (w - p(x))y &= (w - p(x))[(1 - w)y_0 + wy_1] = w(1 - w)y_0 - p(x)(1 - w)y_0 + w^2y_1 - p(x)wy_1 \\ &= wy_0 - w^2y_0 - p(x)(1 - w)y_0 + w^2y_1 - p(x)wy_1 \end{aligned} \quad (37)$$

Como w é uma variável binária que só assume valores 0 e 1, vale que $w = w^2$, então

$$(w - p(x))y = wy_1 - p(x)(1 - w)y_0 - p(x)wy_1 \quad (38)$$

A seguir, extrai-se dos dois lados da equação o valor esperado condicional a w e x :

$$E[(w - p(x))y|w, x] = E[wy_1|w, x] - E[p(x)(1 - w)y_0|w, x] - E[p(x)wy_1|w, x] \quad (39)$$

Como vale a Hipótese 2:

$$E[(w - p(x))y|x] = wE[y_1|x] - p(x)(1 - w)E[y_0|x] - p(x)wE[y_1|x] \quad (40)$$

Em seguida, aplica-se novamente o valor esperado condicional, dessa vez apenas a x . Por definição, $E[w|x] = p(x)$. O lado esquerdo da igualdade não se altera. Temos então:

$$\begin{aligned} E[(w - p(x))y|x] &= p(x)E[y_1|x] - p(x)(1 - p(x))E[y_0|x] - (p(x))^2 E[y_1|x] \\ &= p(x)[1 - p(x)]E[y_1|x] - p(x)[1 - p(x)]E[y_0|x] \end{aligned} \quad (41)$$

Agrupando os termos, ficamos com:

$$E[(w - p(x))y|x] = p(x)[1 - p(x)] E[y_1 - y_0|x] \quad (42)$$

Reescrevendo:

$$E[y_1 - y_0|x] = \frac{E[(w - p(x))y|x]}{p(x)[1 - p(x)]} \quad (43)$$

O lado esquerdo desta igualdade é o ATE condicional a x . Pela lei das expectativas iteradas, pode-se então escrever:

$$E[y_1 - y_0] = ATE = E \left[\frac{(w - p(x))y}{p(x)(1 - p(x))} \right] \quad (44)$$

Está provado, portanto, que o ATE pode ser escrito como uma função do escore de propensão $p(x)$.

Para estimar o ATE, deve-se, primeiramente, estimar o escore de propensão através do modelo logit que resulta em $\hat{p}(x)$. A seguir, estima-se o valor esperado pela média amostral. Algebricamente:

$$\hat{ATE} = \frac{1}{N} \sum_{i=1}^N \frac{(w_i - \hat{p}_i(x))y_i}{\hat{p}_i(x)(1 - \hat{p}_i(x))} \quad (45)$$

2.3.2. Estimação via regressão

É possível também estimar o ATE a partir de uma regressão linear da variável resposta y na variável tratamento w e na estimativa do escore de propensão $\hat{p}(x)$:

$$y = \beta_0 + \beta_1 w + \beta_2 \hat{p}(x) + u \quad (46)$$

Na equação (46), o coeficiente que multiplica w é o efeito do tratamento. Dependendo das ferramentas à disposição do usuário, este cálculo pode ser mais simples que a estimativa anterior a partir de uma função direta de $\hat{p}(x)$.

Além disso, esta regressão possui um significado intuitivo muito forte baseado na ideia de pareamento de indivíduos. Sabe-se que, em uma regressão linear, um coeficiente que multiplica uma variável representa numericamente o efeito esperado na variável resposta causado pelo aumento de uma unidade na variável em questão se nada mais for alterado, ou, em outras palavras, o efeito *ceteris paribus* do aumento de uma unidade na variável independente em questão. Para mais informações sobre as propriedades e significados básicos dos coeficientes de uma regressão linear, o leitor pode consultar Wooldridge (2000).

Na regressão representada pela equação (46), o coeficiente β_1 representa, portanto, o efeito *ceteris paribus* na variável resposta y de aumentar w em uma unidade. Como, por definição, w só assume valores iguais a 0 e 1, aumentar w em uma unidade significa necessariamente passar de $w = 0$ para $w = 1$. Mas a diferença na variável resposta entre a situação em que $w = 1$ e a situação em que $w = 0$ é exatamente o efeito do tratamento que queremos estimar.

$$ATE = E[y_1 - y_0] = E[y|w = 1] - E[y|w = 0] = E[\beta_0 + \beta_1 + \beta_2\hat{p}(x) + u] - E[\beta_0 + \beta_2\hat{p}(x) + u] \quad (47)$$

$$= E[\beta_1] = \beta_1$$

Portanto, podemos estimar o valor do ATE através da estimação do coeficiente β_1 na equação (46):

$$\hat{ATE} = \hat{\beta}_1 \quad (48)$$

Não será demonstrado neste trabalho, mas a consistência do estimador $\hat{\beta}_1$ se verifica se a seguinte hipótese for verificada:

Hipótese 4: $E[y_1 - y_0|x]$ é descorrelacionado de $Var(w|x) = p(x)(1 - p(x))$

A Hipótese 4 não é muito forte e, em geral, pode ser assumida sem influência significativa no resultado. Isto se deve ao fato de que o valor esperado condicional a x da diferença entre as respostas y_1 e y_0 tem, muito possivelmente, uma relação linear com diversos elementos do vetor x , enquanto a variância de w condicional a x será, muito provavelmente, uma função quadrática em x . Como a correlação é, por definição, uma medida apenas da relação **linear** entre duas variáveis, é plausível que a Hipótese 4 seja verificada. Wooldrige (2010) usa como analogia a esta explicação o fato de que a correlação entre uma variável qualquer a e o seu quadrado a^2 é nula, embora haja uma relação quadrática perfeita entre elas.

Mesmo assim, uma maneira de relaxar a Hipótese 4 é incluir interações entre w e $p(x)$ na regressão.

Primeiro, é necessário antes mostrar que, dado $p(x)$, w é independente do par (y_0, y_1) . Pela lei das expectativas iteradas, pode-se escrever:

$$E[w|y_0, y_1, p(x)] = E[E[w|y_0, y_1, x]|y_0, y_1, p(x)] = E[p(x)|y_0, y_1, p(x)] = p(x) = E[w|p(x)] \quad (49)$$

A partir das equações (2), (9) e (10) podemos escrever:

$$y = y_0 + w(E[y_1] - E[y_0]) + w(v_1 - v_0) \quad (50)$$

Como já foi mostrada a independência entre w e (y_0, y_1) dado $p(x)$, podemos aplicar o valor esperado condicional a w e $p(x)$:

$$E[y|w, p(x)] = E[y_0|p(x)] + w(E[y_1] - E[y_0]) + w(E[v_1 - v_0|p(x)]) \quad (51)$$

Finalmente, podemos repetir o raciocínio do final da seção 2.2 e reescrever v_0 e v_1 como um termo constante e um termo linear em x :

$$E[v_0|x] = \eta_0 + \beta_0 x \quad (52)$$

$$E[v_1|x] = \eta_0 + \beta_1 x \quad (53)$$

Assim, é possível também reescrever a equação acima como uma soma entre um termo constante, um termo função de w , outro termo função de $p(x)$ e um quarto termo função da interação entre w e $p(x)$. Também analogamente à seção 2.2, o vetor $p(x)$ é, sem perda, subtraído de sua média amostral $\overline{p(x)}$. A regressão final fica:

$$E[y|w, p(x)] = \delta_0 + \delta_1 w + \delta_2 p(x) + \delta_3 w(p(x) - \overline{p(x)}) \quad (54)$$

Utilizando a estimativa do escore de propensão calculada através do modelo logit $\hat{p}(x)$, a estimativa do ATE como função de x é, então:

$$\widehat{ATE}(x) = \hat{\delta}_1 + \hat{\delta}_3(\hat{p}(x) - \overline{\hat{p}(x)}) \quad (55)$$

Analogamente ao método de regressão com interações do final da seção 2.2, é possível calcular o efeito do tratamento para qualquer subgrupo de interesse da população fazendo uma média apenas com os valores de $p(x)$ dos indivíduos de interesse. A estimativa em toda a população é:

$$\widehat{ATE} = \hat{\delta}_1 \quad (56)$$

2.4. Estimação através do método de Variáveis Instrumentais

Um modelo de regressão linear em sua forma geral busca estimar o efeito que uma ou mais variáveis independentes têm sobre uma dada variável dependente e pode ser escrito da seguinte forma:

$$y = \beta_0 + \beta x + u \quad (57)$$

Onde x é o vetor com as variáveis explicativas (regressores) e β o vetor de coeficientes que quantifica o efeito *ceteris paribus* de cada uma delas na variável dependente.

Em toda regressão linear, assume-se a hipótese básica de que todas as variáveis explicativas presentes em x são descorrelacionadas do erro u . Em um modelo bem formulado, as variáveis explicativas escolhidas são aquelas que de fato podem ser consideradas como determinantes do resultado y . Desta forma, pode-se considerar como válida a hipótese de inexistência de correlação entre os regressores e u . Se vale esta hipótese, os estimadores para os coeficientes em β e para o intercepto β_0 são consistentes e não-viesados. Para a demonstração da consistência e da ausência de viés dos estimadores dos coeficientes, o leitor pode consultar Wooldridge (2000, páginas 46 a 58)

No entanto, segundo Stock e Watson (2003), um modelo de regressão pode apresentar problemas “de validade interna” que comprometem a consistência dos estimadores dos coeficientes, introduzem um viés indesejado e, em última análise, podem invalidar toda a construção do modelo. Um dos problemas mais comuns desta categoria é o chamado **viés de omissão**, que ocorre quando uma variável explicativa, que teoricamente deve estar presente no modelo, por alguma razão não é incluída.

Se ocorre viés de omissão, o termo de erro u inclui não apenas variações intrínsecas ao modelo, mas também informação acerca da variável que foi indevidamente omitida. Naturalmente, é possível (e até provável) que esta variável omitida seja correlacionada com um ou mais regressores em x . Consequentemente, o erro passa a ser correlacionado com regressores de x e deixa de valer a hipótese fundamental mencionada no início desta seção. Se a variável omitida tiver sido observada, uma maneira simples de corrigir o viés de omissão é, naturalmente, acrescentá-la ao vetor de regressores, garantindo novamente a ausência de correlação com o erro. No entanto, o mais comum é que o viés de omissão seja causado por variáveis não observadas, sendo necessário um novo método para contornar este problema.

2.4.1. Introdução à regressão com Variáveis Instrumentais

O método de regressão através de variáveis instrumentais permite, sob dadas circunstâncias, estimar consistentemente os coeficientes de um modelo em que um ou mais dos regressores é correlacionado com erro u . Primeiro, é interessante reescrever o modelo de regressão explicitando a variável presente no vetor x que é problemática, ou seja, que possui correlação com o erro:

$$y = \beta_0 + \rho w + \beta x + u \quad (58)$$

Onde

$$\text{corr}(w, u) \neq 0$$

Por simplicidade de notação, chamaremos a variável problemática de w pois, mais à frente, na seção 2.4.2, este método será aplicado exatamente supondo que o problema de estimação do ATE está na não aleatoriedade da atribuição do tratamento ao qual nos referimos nas seções anteriores pela letra w . Veremos que este método permite relaxar todas as hipóteses de aleatoriedade feitas até agora, sendo muito útil para a estimativa do ATE.

Por definição, uma variável correlacionada com o erro e , portanto, problemática, é dita **endógena** enquanto uma variável que não se correlaciona com o erro é chamada de **exógena**. A intuição por trás desta nomenclatura se deve ao fato de que variáveis não correlacionadas com u são determinadas unicamente por fatores observados externos ao modelo, enquanto variáveis que se correlacionam com o erro são parcialmente determinadas pelo termo u , não sendo, portanto, totalmente externas ao modelo.

Stock e Watson (2003) propõem uma maneira de pensar o método de variáveis instrumentais imaginando a variável w como tendo duas partes: uma parte que, por alguma razão, possui correlação com o termo de erro e que, portanto, causa o problema no modelo original (endógena) e outra parte que é descorrelacionada de u (exógena).

Por simplicidade, vamos assumir a existência de uma única variável endógena. O método de variáveis instrumentais utiliza uma outra variável, chamada de **instrumento**, para isolar a parte de w que é descorrelacionada com o erro da parte problemática, permitindo então uma estimativa consistente dos coeficientes da regressão.

Chamaremos a variável utilizada como instrumento de z . Para que z seja um instrumento válido, é necessário que sejam respeitados os dois itens da Hipótese 5 apresentada a seguir:

Hipótese 5: existe uma variável z observada na população e ausente no modelo original tal que:

$$(a) \text{corr}(z, w) \neq 0$$

$$(b) \text{corr}(z, u) = 0$$

Em outras palavras, é preciso que haja uma variável observada z que não tenha sido considerada no modelo e que seja correlacionada à variável problemática (respeitando (a)) e exógena ao modelo (respeitando (b)). A generalização para o caso com mais de uma variável endógena é trivial, devendo haver um instrumento z para cada variável endógena no modelo original.

Se z satisfaz a Hipótese 5, o coeficiente ρ da equação (58) pode ser consistentemente estimado através do método de **Mínimos Quadrado em Dois Estágios (MQ2E)**. Este método consiste, como sugere o nome, de duas regressões interligadas através do método de Mínimos Quadrados Ordinários. Para definições e propriedades básicas do estimador de Mínimos Quadrados Ordinários, o leitor pode consultar Wooldridge (2000, capítulos 2, 3 e 5).

O primeiro estágio do método de MQ2E decompõe a variável problemática w em seus componentes endógeno e exógeno e o segundo estima o coeficiente desejado ρ a partir da parte exógena de w .

O primeiro estágio do método de MQ2E decompõe a variável problemática w em suas partes exógena e endógena através da regressão de w no instrumento z :

$$w = \pi_0 + \pi_1 z + \pi x + v \quad (59)$$

Como a variável instrumental z e todas as outras variáveis independentes presentes em x são exógenas ao modelo original (equação (58)), a componente de w que pode ser explicada por x e z , por construção, também é exógena. Ou seja, o termo $\pi_0 + \pi_1 z + \pi x$ é descorrelacionado do erro u da

equação (58). Esta parte da equação (59) representa, portanto, a parte não-problemática de w . Analogamente, v representa a parte de w que não é explicada por z e x e, conseqüentemente, que é endógena ao modelo original.

Pode-se, então, estimar através de mínimos quadrados ordinários os coeficientes da parte exógena de w :

$$\hat{w} = \hat{\pi}_0 + \hat{\pi}_1 z + \hat{\pi} x \quad (60)$$

A variável $\hat{w}$ representa a estimação da variável w expurgada de sua componente endógena, restando apenas a componente exógena e, portanto, não problemática. Para finalizar o método de MQ2E, realiza-se a regressão de y em $\hat{w}$ e nas demais variáveis exógenas presentes no vetor x :

$$y = \beta_0' + \rho' \hat{w} + \beta' x + u \quad (61)$$

Agora, não há mais variáveis endógenas no modelo de regressão para y e, portanto, os estimadores dos coeficientes são consistentes e não-viesados.

Para enriquecer o formalismo da metodologia apresentada, é pertinente deduzir a fórmula do estimador para ρ' no caso mais simples em que w é a única variável do modelo de regressão, não havendo outras variáveis independentes em x . Pela fórmula do estimador do coeficiente linear em uma regressão simples:

$$\hat{\rho}' = \frac{S_{\hat{w}y}}{S_{\hat{w}}^2} \quad (62)$$

Onde $S_{\hat{w}y}$ é a covariância amostral entre $\hat{w}$ e y e $S_{\hat{w}}^2$ é a variância amostral de $\hat{w}$.

Mas sabemos que:

$$\hat{w} = \hat{\pi}_0 + \hat{\pi}_1 z \quad (63)$$

Então, pela definição de covariância amostral:

$$S_{\hat{w}y} = \hat{\pi}_1 S_{zy} \quad (64)$$

E pela definição de variância amostral:

$$S_{\hat{w}}^2 = \hat{\pi}_1^2 S_z^2 \quad (65)$$

Logo:

$$\hat{\rho}' = \frac{\hat{\pi}_1 S_{zy}}{\hat{\pi}_1^2 S_z^2} = \frac{S_{zy}}{\hat{\pi}_1 S_z^2} \quad (66)$$

Mas $\hat{\pi}_1$ é o estimador do coeficiente linear da regressão de w em z , portanto:

$$\hat{\pi}_1 = \frac{S_{zw}}{S_z^2} \quad (67)$$

Logo, substituindo a equação (67) na equação (66), obtemos:

$$\hat{\rho}' = \frac{S_{zy}}{S_{zw}} \quad (68)$$

Para provar a consistência de $\hat{\rho}'$, mostraremos sua convergência em probabilidade para o coeficiente populacional da regressão através do operador matemático $plim()$.

De Angrist e Pischke (2009, páginas 34 a 42), podemos escrever o coeficiente populacional ρ' da regressão de y em w (que é, portanto, o coeficiente que deseja-se estimar) como:

$$\rho' = \frac{\text{cov}(y, w)}{\text{var}(w)} = \frac{E(wy) - E(w)E(y)}{E(w^2) - E(w)^2} \quad (69)$$

Como a covariância amostral S é um estimador consistente da covariância populacional, podemos usar as propriedades do operador $\text{plim}()$ e escrever:

$$\text{plim}(\widehat{\rho'}) = \frac{\text{plim}(S_{zy})}{\text{plim}(S_{zw})} = \frac{\text{cov}(z, y)}{\text{cov}(z, w)} = \frac{E(zy) - E(z)E(y)}{E(zw) - E(z)E(w)} \quad (70)$$

Podemos a seguir, reescrever a equação (59) (sem o vetor x , pois estamos tratando do caso simplificado em que w é o único regressor) colocando z em evidência:

$$z = \frac{w}{\pi_1} - \frac{\pi_0}{\pi_1} - \frac{v}{\pi_1} \quad (71)$$

A seguir, a partir da equação (71), podemos tirar o valor esperado de z , zy e zw :

$$E(z) = E\left(\frac{w}{\pi_1} - \frac{\pi_0}{\pi_1} - \frac{v}{\pi_1}\right) = \frac{E(w)}{\pi_1} - \frac{\pi_0}{\pi_1} \quad (72)$$

$$E(zy) = E\left(\frac{wy}{\pi_1} - \frac{\pi_0 y}{\pi_1} - \frac{vy}{\pi_1}\right) = \frac{E(wy)}{\pi_1} - \frac{\pi_0}{\pi_1} E(y) \quad (73)$$

$$E(zw) = E\left(\frac{w^2}{\pi_1} - \frac{\pi_0 w}{\pi_1} - \frac{vw}{\pi_1}\right) = \frac{E(w^2)}{\pi_1} - \frac{\pi_0}{\pi_1} E(w) \quad (74)$$

Finalmente, substituindo as equações (72), (73) e (74) na equação (70) e multiplicando o numerador e denominador por π_1 , obtêm-se:

$$\text{plim}(\widehat{\rho'}) = \frac{E(zy) - E(z)E(y)}{E(zw) - E(z)E(w)} = \frac{E(wy) - \pi_0 E(y) - E(w)E(y) + \pi_0 E(y)}{E(w^2) - \pi_0 E(w) - E(w)E(w) + \pi_0 E(w)} = \frac{E(wy) - E(w)E(y)}{E(w^2) - E(w)^2} \quad (75)$$

Portanto:

$$\text{plim}(\widehat{\rho'}) = \rho' \quad (76)$$

Ou seja, o estimador $\widehat{\rho'}$ tende em probabilidade ao parâmetro populacional ρ' e, portanto, é consistente.

Além disso, $\widehat{\rho'}$ é também normalmente distribuído para amostras grandes. Esta propriedade se deve ao fato de que, pela definição de covariância amostral, o estimador pode ser escrito a partir de médias amostrais de variáveis. Pelo Teorema Central do Limite, para amostras muito grandes, médias amostrais de variáveis seguem uma distribuição Normal. Em outras palavras, $\widehat{\rho'}$ é assintoticamente normal (Stock e Watson (2003)).

Angrist e Pischke (2009) mostram que, no caso mais geral, com outras variáveis independentes no modelo da equação (59) além de w , ou seja, quando consideramos uma regressão múltipla incluindo os regressores no vetor x , o estimador pode ser escrito como:

$$\widehat{\rho'} = \frac{S_{y\widehat{w}}}{S_{w\widehat{w}}} \quad (77)$$

Onde $\widehat{w}$ é uma variável obtida através dos resíduos da regressão de $\widehat{w}$ nos demais regressores exógenos presentes em x .

O método de regressão através de variáveis instrumentais é uma ferramenta poderosa que permite corrigir a endogeneidade de variáveis explicativas, que é um dos principais problemas econométricos quando se trabalha com dados reais. No entanto, para que ele funcione, é imprescindível a existência de um instrumento adequado que respeite integralmente as condições da Hipótese 5. Embora muitas vezes, em bancos de dados completos e variados, isto seja possível, a obtenção de um instrumento nem sempre é trivial e o uso de instrumentos inadequados invalida a consistência e introduz um viés nas estimativas de todos os coeficientes.

A Tabela 1 resume os passos necessários à estimação de um modelo de regressão através da metodologia de variáveis instrumentais.

Passo	Procedimento
1	Escolha de uma variável instrumental z que respeite a hipótese 5 (descorrelatada do erro e correlatada com o regressor endógeno w)
2	Regressão de w em z e x (primeiro estágio do MQ2E)
3	Regressão de y em $\hat{w}$, proveniente da estimação do passo anterior e x (segundo estágio do MQ2E)

Tabela 1 - Passo-a-passo para a estimação de um modelo de regressão com variáveis instrumentais

A seguir, veremos com este método pode ser aplicado ao cálculo do Efeito Médio do Tratamento (ATE).

2.4.2. Aplicação à estimação do ATE

Como foi dito na seção 2.2, a Hipótese 2, que foi assumida em todas as metodologias de estimação do ATE até aqui vistas, nem sempre é verificada na prática e pode acabar sendo muito restritiva para trabalhos econométricos reais, em que limitações práticas e observacionais devem ser levadas em conta. É possível, por exemplo, em uma situação em que a atribuição do tratamento w não foi aleatória, que não seja possível observar variáveis independentes o suficiente para explicar a atribuição do tratamento. Nesta situação, a Hipótese 2 é violada. O método de variáveis instrumentais descrito em 2.4.1 pode ser uma solução adequada para esses casos.

Violar a Hipótese 2 significa dizer que na regressão da variável resposta y no tratamento w e nas demais variáveis explicativas x , existe viés causado pela não-aleatoriedade da atribuição de w que não é explicado pelos demais regressores em x . Essencialmente, isto significa dizer que na regressão de y em w e x , existe correlação entre w e o termo de erro u . Algebricamente:

$$y = \beta_0 + \rho w + \beta x + u \quad (78)$$

$$\text{corr}(w, u) \neq 0$$

Se a Hipótese 2 é violada, o estimador desta regressão será inconsistente. Na dedução dos estimadores do ATE através do escore de propensão, também assumimos a validade da Hipótese 2. Desta maneira, nem a metodologia de estimação do ATE através de um modelo de regressão na seção 2.2, nem sua estimação através do escore de propensão na seção 2.3 serão válidas.

Para contornar esta limitação, deve-se utilizar um instrumento adequado para o tratamento respeitando a Hipótese 5. Relembrando, seja z a variável instrumental para o tratamento w , deve-se garantir que:

$$(a) \text{corr}(z, w) \neq 0$$

$$(b) \text{corr}(z, u) = 0$$

Ou seja, é necessário que a variável observada z seja correlacionada à atribuição do tratamento, mas exógena ao modelo de regressão representado pela equação (78). Stock e Watson (2003, páginas 359 a 365) argumentam que, muitas vezes, uma variável que seja atribuída de maneira aleatória à população é, por construção, desconhecida do termo de erro u . Se esta variável tiver alguma influência prática na atribuição do tratamento, trata-se de um instrumento apropriado.

Wooldridge (2010, página 622) exemplifica que, muitas vezes, embora a atribuição do tratamento não seja feita de maneira aleatória, a elegibilidade de um indivíduo para ser submetido pode ser aleatória e, dessa maneira, um instrumento eficiente para o tratamento. Em um programa de treinamento para um dado emprego, por exemplo, vários indivíduos podem ser aleatoriamente convidados para participar. Neste exemplo, a participação em si no programa de treinamento não é aleatória, pois depende da decisão de cada indivíduo de aceitar ou não o convite. No entanto, existe uma óbvia correlação entre o convite para participar e a participação em si, pois só aqueles que foram convidados podem tomar parte no programa. Neste exemplo, a variável binária que denota se um indivíduo foi ou não convidado para o programa pode ser um instrumento eficiente para a variável binária que indica a participação de fato.

Uma vez encontrado um instrumento z apropriado, o procedimento de MQ2E deve ser seguido tal como apresentado na seção 2.4.1. Primeiro, realiza-se uma regressão da variável tratamento na variável instrumental e nas demais variáveis independentes presentes no vetor x , encontrando assim uma estimativa para o tratamento $\hat{w}$:

$$\hat{w} = \hat{\pi}_0 + \hat{\pi}_1 z + \hat{\pi} x \quad (79)$$

Expurgado o componente endógeno do tratamento, fazemos a regressão da variável resposta y em $\hat{w}$ e nas variáveis presentes em x :

$$y = \beta_0' + \rho' \hat{w} + \beta' x + u \quad (80)$$

Novamente, como a equação (80) foi construída a partir de variáveis exógenas, o estimador para o coeficiente que multiplica a variável $\hat{w}$ é consistente e não-viesado, e, portanto:

$$ATE = \hat{\rho}' \quad (.81)$$

Finalmente, cabe ainda uma observação importante acerca do primeiro estágio do método de MQ2E. Como a variável w só assume valores 0 e 1, o leitor pode se perguntar porque não estimamos $\hat{w}$ por um modelo mais adequado para estimação de uma variável binária como, por exemplo, um modelo probit ou logit, tal como o usado para estimar $p(x)$ na seção 2.2. O problema está no fato de que modelos logit e probit são, por construção, altamente não lineares nos regressores. No entanto, o primeiro estágio do método MQ2E (equação (59)) só produz resíduos que são realmente não correlacionados com os regressores se a estimação for feita através de Mínimos Quadrados Ordinários, o que não ocorre em modelos não lineares como probit e logit. Além disso, em nenhum momento foi feita referência à qualidade da estimação de w através de $\hat{w}$. De fato, não é necessário que $\hat{w}$ aproxime corretamente a variável original w , apenas que o expurgue de sua componente endógena. Não é, portanto, um problema no método MQ2E se a variável $\hat{w}$ assumir, por exemplo, valores maiores que 1 ou menores do que 0, pois estes valores não serão usados diretamente para estimar ou prever o tratamento, mas apenas como um passo intermediário no método de MQ2E. Desta forma, uma estimação através de um modelo linear como na equação (60) é perfeitamente adequada para o método de variáveis instrumentais. Nas palavras de Angrist e Pischke (2009): "Por que correr riscos com um modelo não-linear se não precisamos?"

2.5. Resumo das metodologias de estimação apresentadas

A Tabela 2 resume as sete metodologias de estimação do ATE desenvolvidas até aqui neste trabalho.

	Método	Hipóteses	Procedimento para estimação do ATE
1	Diferença simples das médias dos grupos	1	$\hat{ATE} = \bar{y}_1 - \bar{y}_0$
2	Regressão direta sem termo de interação	2 e 3	Regressão de y em w e x
3	Regressão direta com termo de interação	2	Regressão de y em w, x e $(x - \bar{x})w$
4	Função do Escore de Propensão	2	$\hat{ATE} = \frac{1}{N} \sum_{i=1}^N \frac{(w_i - \hat{p}_i(x))y_i}{\hat{p}_i(x)(1 - \hat{p}_i(x))}$
5	Regressão com Escore de Propensão sem termo de interação	2 e 4	Regressão de y em w e $\hat{p}(x)$
6	Regressão com Escore de Propensão com termo de interação	2	Regressão de y em w e $\hat{p}(x)$ e $w(\hat{p}(x) - \bar{\hat{p}})$
7	Variáveis Instrumentais	5	MQ2E: regressão de y em $\hat{w}$ e x , onde $\hat{w}$ é a estimativa de w a partir de sua regressão em z e x

Tabela 2 - Resumo das metodologias de estimação do ATE apresentadas

O método 1, descrito na seção 2.1.2, é o mais simples, mas também o mais restritivo de todos. A Hipótese 1 assume total independência entre a atribuição do tratamento e o resultado, o que em geral não é verdade. Este método só pode ser aplicado eficientemente em experimentos aleatorizados controlados. O ATE é apenas a diferença das médias entre o grupo tratado e o grupo não tratado.

Os métodos 2 e 3, descritos na seção 2.2, são baseados em modelos de regressão linear múltipla com a variável tratamento como um dos regressores. O nível de restritividade da Hipótese 2 depende do banco de dados estudado. Se há variáveis suficientes para explicar a atribuição do tratamento, a Hipótese 2 se verifica e o uso destes métodos é eficiente. A diferença entre os dois métodos é a inclusão do termo de interação entre o tratamento e os demais regressores. Como a Hipótese 3 não é muito restritiva e, em geral, se verifica, os métodos 2 e 3 tendem a produzir resultados semelhantes.

O método 4, descrito na seção 2.3.1, estima o ATE como uma função direta do escore de propensão. Depende da mesma Hipótese 2, cujo nível de restritividade depende do banco de dados estudado.

Os métodos 5 e 6, descritos na seção 2.3.2 e também baseados no escore de propensão, partem do princípio de que o coeficiente da regressão que multiplica a variável tratamento representa o efeito *ceteris paribus* do efeito do tratamento. Também depende da validade da Hipótese 2 que está atrelada a características do banco de dados e da possibilidade de estimação de um modelo logit. Como a Hipótese 4 não é muito restritiva e, em geral, se verifica, os métodos 5 e 6 tendem a produzir resultados semelhantes.

O método 7, descritos na seção 2.4, é o mais geral e mais eficiente de todos e tende a produzir os resultados mais corretos, visto que a Hipótese 5 é testada de maneira empírica com facilidade e, caso exista um instrumento adequado, a estimação é consistente e não-viesada. É o método ideal para bancos de dados nos quais os outros métodos não produzem bons resultados, em geral por violação da Hipótese 2. Sua dificuldade reside mais no entendimento da metodologia, que depende de uma teoria

um pouco mais avançada de econometria e do uso de um bom instrumento, do que em empecilhos operacionais.

Como observado no início deste capítulo, existe um compromisso entre a complexidade metodológica do método e o quão restritivas são as hipóteses necessárias para aplicá-lo. Cada estudo possui dados e características muito próprios e, portanto, pode ser mais bem realizado com um metodologia diferente. Ainda assim, de modo geral, o método 7, baseado em Variáveis Instrumentais, tende a ser o mais completo e, quando possível, deve ser priorizado.

3. Estudo de Caso

3.1. O Projeto Tennessee STAR

Muito se discute no ramo de planejamento de políticas educacionais a respeito dos efeitos do tamanho das turmas na qualidade do aprendizado dos alunos, principalmente nos primeiros anos escolares. Muitos defendem que turmas menores aumentam o contato aluno-professor e melhoram o aprendizado e o desempenho dos alunos. Entretanto, a comprovação empírica desta tese exige estudos vastos e, muitas vezes, impraticáveis por questões econômicas e operacionais.

Nos anos 80, o primeiro experimento em larga escala visando uma análise quantitativa deste tipo de política foi conduzido no estado americano do Tennessee pelas autoridades estaduais. O projeto foi batizado de *Tennessee Student/Teacher Achievement Ratio* (em uma tradução livre, razão de desempenho aluno/professor do Tennessee), sendo referido mais geralmente por sua abreviação: Tennessee STAR.

A primeira importante análise econométrica dos dados resultantes desse experimento foi o artigo de Krueger (1997). Angrist e Pischke (2009) e Stock e Watson (2003) também desenvolveram análises a respeito dos dados do projeto e chegaram a resultados interessantes e bastante semelhantes. Esta seção é fortemente baseada em informações extraídas destas três referências.

Custeado pelo governo estadual, o projeto STAR foi orçado em aproximadamente 12 milhões de dólares (valor em dólares de 1985, o que equivaleria a cerca de 26 milhões de dólares de 2015) e envolveu um total de aproximadamente 11600 alunos de 80 escolas do estado do Tennessee. O experimento durou quatro anos e acompanhou alunos durante o jardim de infância (que, nos Estados Unidos, é a série imediatamente antes do início da escola regular), a primeira, a segunda e a terceira série primárias.

No ano de 1985, os alunos do jardim de infância das escolas estaduais participantes no experimento foram aleatoriamente divididos em três grupos para comparação de resultados. O primeiro e o segundo grupos de alunos foram mantidos em turmas de tamanho regular (22 a 25 alunos por turma), sendo que, para as turmas do segundo grupo, um professor assistente trabalhou em tempo integral auxiliando o professor principal. Os alunos do terceiro grupo foram deslocados para turmas de tamanho reduzido (13-17 alunos por turma). Os alunos, que deveriam se manter nas mesmas turmas a que foram inicialmente atribuídos no jardim de infância pelos quatro anos seguintes, foram submetidos a testes padronizados de desempenho em inglês e matemática ao final do jardim de infância e da primeira, segunda e terceira séries. Desta maneira, o efeito da inclusão de um assistente e da redução do tamanho da turma seriam medidos a curto prazo, nos resultados dos testes aplicado ao final do próprio jardim de infância, e também a médio e longo prazo com a análise dos resultados dos testes aplicados nos anos subsequentes à divisão inicial dos alunos participantes.

Inicialmente, conforme já foi mencionado, pretendia-se que o tipo de turma ao qual cada aluno foi atribuído fosse inteiramente aleatório (e o mesmo) durante todos os quatro anos em que os dados foram captados. Desta maneira, a análise dos resultados seria mais confiável e simples (como visto na seção 2.1.2, experimentos aleatorizados tem sua análise operacionalmente facilitada).

No entanto, ao longo dos quatro anos, houve considerável violação desta aleatoriedade inicialmente planejada. Como era de se esperar, várias crianças novas matricularam-se nas escolas participantes no decorrer do experimento e tiveram que ser alocadas em turmas previamente divididas. Outras tantas mudaram-se de estado e deixaram o experimento antes de seu término. Muitas crianças tiveram que ser realocadas e mudaram de turma dentro das escolas por questões comportamentais e disciplinares. Finalmente, diversos pais adeptos da ideia de que turmas pequenas influem positivamente no aprendizado das crianças pediram às escolas que deslocassem seus filhos para as turmas pequenas no decorrer do estudo. Este último talvez tenha sido o principal fator causador da quebra de aleatoriedade

da seleção original. Ao final dos quatro anos, portanto, uma parcela significativa dos alunos, agora já na terceira série primária, não estava mais em sua turma de origem.

Tais fatos, como dito, contribuíram não só para uma alteração no tamanho das turmas inicialmente planejado, mas, principalmente, para uma gradual quebra na aleatoriedade inicial da atribuição de turmas. Os motivos que levam, por exemplo, pais a pressionarem a escola por uma mudança de turma de seu filho são certamente não aleatórios e também de difícil quantificação, o que introduz vieses relacionados à seleção do tratamento e à omissão de possíveis variáveis explicativas associadas. Desta maneira, as técnicas de análise econométrica dos dados do estudo devem levar em consideração uma seleção não aleatória do tipo de turma ao qual cada aluno foi atribuído.

O objetivo deste trabalho é realizar um estudo de caso aplicando as técnicas de estimação de efeito médio do tratamento (ATE) ao experimento Tennessee STAR e tecer conclusões a respeito dos resultados encontrados

3.2. Descrição do banco de dados

As subseções seguintes apresentam uma descrição do significado de cada variável utilizada no estudo e uma tabela com estatísticas descritivas relevantes das mesmas. O banco de dados original está disponível ao público em <http://economics.mit.edu/faculty/angrist/data1/mhe/krueger>, uma página aberta com informações adicionais ao texto de Angrist e Pischke (2009). A base original foi submetida a um processo de pré-tratamento visando melhor atender aos objetivos deste trabalho.

Os diversos estudos econométricos realizados com os dados do projeto STAR (dentre os quais podemos destacar Krueger (1997), Angrist e Pischke (2009) e Stock e Watson (2003)) concluíram que a atribuição de um professor assistente para auxiliar o professor principal não tem qualquer influência significativa nos resultados dos testes. Como o objetivo deste estudo de caso é quantificar o efeito da redução do tamanho da turma no desempenho dos alunos, os alunos que foram alocados em turmas de tamanho regular com assistência foram considerados como parte do mesmo grupo que os alunos com turmas de tamanho regular sem assistência. Desta maneira, os três grupos do estudo original tornaram-se apenas dois, entre os quais a única diferença é o tamanho da turma.

Além disso, dada a grande quantidade de alunos envolvidos, permitiu-se aqui analisar apenas aqueles alunos que participaram do estudo em sua totalidade, isto é, permaneceram nas escolas estaduais do Tennessee participantes do experimento entre o jardim de infância e a terceira série. Alunos que abandonaram o experimento ou que se matricularam no decorrer dele foram excluídos do banco de dados.

As unidades observacionais são, portanto, os 3005 alunos que participaram integralmente do estudo. Destes, 436 (14.8%) cursaram a terceira série em um tipo de turma diferente daquela que lhe foi aleatoriamente atribuída no jardim de infância.

O objetivo deste estudo é utilizar as técnicas de estimação do efeito da turma para casos em que a atribuição de tratamento não foi aleatória. Conforme mencionado na seção anterior, a quebra de aleatoriedade deu-se gradualmente ao longo dos quatro anos de estudo. Assim sendo, a atribuição de turmas ao final do estudo, isto é, na terceira série primária, era a que continha a maior quebra na aleatoriedade inicial.

Desta maneira, a variável tratamento (w) utilizada no estudo de caso é o tipo de turma (regular ou de tamanho reduzido) em que cada estudante estava na terceira série primária. Portanto, a variável resultado (y) na qual deseja-se estimar o efeito do tratamento é o resultado combinado nos testes de inglês e matemática aplicados aos alunos no final da terceira série primária.

Este estudo de caso foi realizado integralmente utilizando o software de análise estatística R, que é aberto e está gratuitamente disponível para download em <http://www.r-project.org/>.

3.2.1. Descrição das variáveis

Após pré-processamento dos dados, as seguintes variáveis foram incluídas no banco com valores observados para os 3005 estudantes do projeto STAR:

- *score* – a **variável resultado** do estudo, representa a pontuação total de cada estudante nos testes padronizados de inglês e matemática aplicados ao final da terceira série primária. Os dois testes, de inglês e matemática, variam em uma escala de 200 a 800 pontos. A variável *score*, utilizada neste estudo, é a soma da pontuação nos dois testes, variando, portanto, entre 400 e 1600.
- *cltype3* – a **variável tratamento** do estudo, trata-se de uma variável binária que vale 1 caso o estudante tenha cursado a terceira série em uma turma de tamanho reduzido e 0 caso ele tenha cursado a terceira série em uma turma de tamanho regular.
- *cltypek* – variável binária que representa a turma para a qual cada aluno foi aleatoriamente selecionado no início do experimento, isto é, no jardim de infância. Vale 1 caso o aluno tenha sido atribuído a uma turma reduzida e 0 caso ele tenha sido atribuído a uma turma de tamanho regular.
- *ssex* – variável binária para o sexo do aluno, igual a 1 se o aluno for do sexo feminino e 0 caso seja do sexo masculino.
- *srace* – variável binária para a etnia do aluno, igual a 1 caso ele seja da etnia branca e 0 caso contrário. Originalmente, esta era uma variável categórica na qual a etnia era discriminada entre os diversos tipos. No entanto, havia muito poucos alunos que não se enquadravam nas categorias “branco” ou “negro”. A variável foi então transformada em uma binária para indicar se aluno foi classificado como “branco” ou não.
- *trace3* – variável binária para a etnia do professor da terceira série primária, igual a 1 caso ele seja da etnia branca e 0 caso contrário. O tratamento dado a esta variável foi análogo à variável *srace*.
- *schtype3* – variável binária para a localização da escola, igual a 1 caso a escola seja localizada em uma zona rural e 0 caso seja localizada em uma zona urbana. Originalmente, tratava-se de uma variável categórica que dividia as escolas urbanas em diversos tipos, tais como suburbana ou central. Após uma análise preliminar dos dados, optou-se por agrupar os diversos tipos de escola urbana e criar uma variável binária.
- *totexp3* – tempo de experiência de ensino, em anos, do professor da terceira série primária.
- *hdeg3* – variável binária para a qualificação acadêmica do professor da terceira série, igual a 1 caso ele tenha algum curso de pós-graduação e 0 caso tenha apenas graduação. Originalmente esta era uma variável categórica que discriminava os diversos tipos de pós-graduação. Como a maioria dos diplomas de pós-graduação eram de mestrado, havendo muito poucos casos dos outros tipos de diploma, optou-se pela criação de uma variável binária.
- *ses3* – variável binária que indica a renda familiar do aluno na terceira série. Nas escolas do Tennessee, todos os alunos são submetidos a uma avaliação sócio-econômica que analisa as condições financeiras de seus pais. Aqueles alunos cuja renda familiar for considerada baixa têm direito a almoçar gratuitamente na escola. Os alunos cuja renda familiar for considerada alta não recebem almoço gratuito. Esta variável binária vale 1 caso o aluno tenha direito ao almoço gratuito (e, portanto, tenha uma renda familiar considerada pelas autoridades como baixa) e 0 caso contrário.

3.2.2. Estatísticas descritivas

A Tabela 3 a seguir apresenta algumas estatísticas descritivas das variáveis listadas na seção anterior. Estão tabelados a média, o desvio padrão e os percentis 5% e 95% para todas as variáveis. Para as variáveis binárias, naturalmente, a média é equivalente à proporção de valores iguais a 1.

Estatísticas Descritivas				
Variável	Média/Pr(1)	Desvio Padrão	P5	P95
<i>score</i>	1251	72,14	1135	1371
<i>cltype3</i>	0,3877	0,4873	0	1
<i>cltypek</i>	0,3118	0,4633	0	1
<i>ssex</i>	0,5171	0,4998	0	1
<i>srace</i>	0,7354	0,4412	0	1
<i>trace3</i>	0,8186	0,3854	0	1
<i>schtype3</i>	0,6409	0,4798	0	1
<i>totexp3</i>	18,65	12,52	3	38
<i>hdeg3</i>	0,4273	0,4948	0	1
<i>ses3</i>	0,3697	0,4828	0	1

Tabela 3 – Média, desvio padrão e percentis 5% e 95% das variáveis do banco de dados do projeto STAR.

Na Tabela 3, observa-se uma diferença entre os valores médios das variáveis *cltypek* e *cltype3*, que representam, respectivamente, a proporção de alunos em turmas pequenas no jardim de infância e na terceira série primária. Se o planejamento original do experimento tivesse sido respeitado, esses dois valores teriam que ser iguais, pois os alunos se manteriam na mesma turma ao longo de todo o estudo.

No início do experimento, 31,18% dos alunos foram atribuídos aleatoriamente a turmas de tamanho reduzido. Na terceira série, porém, o percentual de alunos em turmas reduzidas foi de 38,77% (aumento de 24,34%). Este número mostra claramente a quebra de aleatoriedade que violou os parâmetros originais do experimento com o passar dos anos. Além disso, um percentual significativamente maior de alunos em turmas reduzidas na terceira série sugere um padrão deliberado de mudança, que está de acordo com o fato de que diversos pais, entusiastas da ideia de que turmas menores beneficiam o aprendizado, pediram para que seus filhos fossem trocados de turmas regulares para turmas reduzidas.

No próximo capítulo, as diversas técnicas de estimação do ATE serão aplicadas aos dados do projeto STAR e seus resultados analisados.

Neste capítulo, cada um dos sete métodos de estimação do ATE apresentados no Capítulo 2 e resumidos na Tabela 2 será aplicado ao banco de dados descrito no Capítulo 3 referente aos resultados do projeto educacional Tennessee STAR. O objetivo é estimar o efeito da redução no tamanho das turmas da terceira série primária sobre o resultado em testes padronizados de inglês e matemática.

Em todos os casos, a variável binária referente ao tratamento (chamada no desenvolvimento teórico de w) é a variável *cltype3*, que define o tipo de turma em que cada aluno cursou a terceira série primária (1 para turma reduzida, 0 para turma de tamanho regular). A variável resultado (chamada no desenvolvimento teórico das metodologias de y) é *score*, que quantifica a pontuação combinada de cada aluno nos testes de inglês e matemática aplicados ao final da terceira série primária. As demais variáveis descritas na Seção 3.2.1, com a exceção da variável *cltypek* (tamanho de turma aleatoriamente atribuído no jardim de infância), fazem parte do vetor de variáveis observadas associadas a cada indivíduo (chamado no desenvolvimento teórico de x).

A ordem e a numeração dos métodos de estimação seguem a numeração da Tabela 2. 5% foi o nível de significância usado como referência para os modelos de regressão construídos.

4.1. Método 1

Como visto, o primeiro método de estimação do ATE consiste em assumir como verdadeira a improvável Hipótese 1, que pressupõe total aleatoriedade na atribuição do tratamento. O estimador do ATE é obtido pela simples diferença entre as médias dos grupos tratado e não tratado.

No estudo de caso aqui realizado, este método estima o ATE através da diferença entre a nota média dos alunos que cursaram a terceira série em turmas reduzidas e a nota média dos alunos que cursaram a terceira série em turmas regulares. O resultado, portanto, é:

$$A\hat{T}E_1 = 10,47$$

4.2. Método 2

O método 2 consiste em uma regressão da variável resultado *score* na variável tratamento *cltype3* e nas demais variáveis independentes presentes no vetor x . Como consta na Tabela 2, assume-se serem verdadeiras as Hipótese 2 e 3. Mais a frente, veremos que, nesse caso específico, não é razoável assumir a validade da Hipótese 2.

Primeiramente, realizou-se uma regressão com todas as variáveis observadas presentes em x .

Variável	Coefficiente	p-valor
Intercepto	1230,62	0
<i>cltype3</i>	11,79	3,75E-06
<i>ssex</i>	6,18	0,01
<i>srace</i>	26,98	5,17E-11
<i>trace3</i>	7,1	0,08
<i>schtype3</i>	-2,65	0,44
<i>totexp3</i>	-0,09	0,39
<i>hdeg3</i>	4,87	0,06
<i>ses3</i>	-33,29	2,11E-30
$R^2 = 0,1219$		

Tabela 4 - coeficientes da regressão do resultado *score* no tratamento *cltype3* e em todas as demais variáveis

As variáveis *trace3*, *schtype3* e *totexp3* não são estatisticamente significantes ao nível de 5% pois seu p-valor é maior do que 0,05. Pelo princípio de parcimônia, elas serão então excluídas do modelo. A variável *hdeg3* possui um p-valor muito próximo de 0,05 (ainda que um pouco maior) e, portanto, será mantida junto com as demais, que são estatisticamente significantes a 5%.

Variável	Coefficiente	p-valor
Intercepto	1232,02	0
<i>cltype3</i>	11,68	4,46E-06
<i>ssex</i>	6,08	0,01
<i>srace</i>	28,63	9,46E-20
<i>hdeg3</i>	5,18	0,04
<i>ses3</i>	-33,67	1,66E-31
$R^2 = 0,1229$		

Tabela 5 - coeficientes da regressão do resultado score no tratamento *cltype3* e nas variáveis independentes que são estatisticamente significantes ao nível de 5%

As variáveis presentes no modelo da Tabela 5 serão as escolhidas para serem usadas em todos os demais modelos deste estudo de caso. Agora, todos os coeficientes são significantes a 5%. A estimativa do ATE é o coeficiente da variável tratamento *cltype3*.

$$A\hat{T}E_2 = 11,68$$

4.3. Método 3

O método 3 consiste em relaxar a Hipótese 3 aplicada ao método 2. Para tal, deve-se incluir termos de interação entre o tratamento e as outras variáveis observadas. Para isto, foram criadas as seguintes variáveis adicionais:

- *ssexnorm_int*: variável *ssex*, que representa o sexo do aluno, subtraída de sua média e multiplicada pelo tratamento *cltype3*
- *sracenorm_int*: variável *srace*, que representa a etnia do aluno, subtraída de sua média e multiplicada pelo tratamento *cltype3*
- *hdeg3norm_int*: variável *hdeg3*, que representa o grau de qualificação do professor, subtraída de sua média e multiplicada pelo tratamento *cltype3*
- *ses3norm_int*: variável *ses3*, que representa o status sócio-econômico do aluno, subtraída de sua média e multiplicada pelo tratamento *cltype3*

O modelo agora é a regressão da variável *score* na variável tratamento *cltype3*, nas variáveis estatisticamente significantes obtidas no método 2 e nas variáveis de interação criadas.

Variável	Coefficiente	p-valor
Intercepto	1226,64	0
<i>cltype3</i>	11,55	5,70E-06
<i>ssex</i>	9,36	3,06E-03
<i>srace</i>	31,56	3,26E-15
<i>hdeg3</i>	6,65	0,04
<i>ses3</i>	-31,39	7,61E-18
<i>ssexnorm_int</i>	-8,49	0,09
<i>sracenorm_int</i>	-8,36	0,20
<i>hdeg3norm_int</i>	-4,62	0,38
<i>ses3norm_int</i>	-6,36	0,28
$R^2 = 0,1244$		

Tabela 6 - coeficientes da regressão do resultado *score* no tratamento, nas variáveis estatisticamente significantes a 5% e na interação do tratamento com as demais variáveis

Os resultados apresentados na Tabela 6 mostram que nenhuma das variáveis de interação é estatisticamente significativa a 5%. Além disso, o valor do coeficiente da variável tratamento é muito semelhante ao obtido no método 2. Isto sugere que, uma vez assumida a validade da Hipótese 2, assumir a Hipótese 3 (necessária para o método 2) é pouco restritivo para o modelo.

O estimador do ATE através deste método é, então:

$$A\hat{T}E_3 = 11,55$$

4.4. Método 4

O método 4, que também assume validade da Hipótese 2, consiste na estimação do ATE como uma função direta de uma estimativa do escore de propensão. Para tal, primeiro realiza-se uma estimação do escore de propensão através, como já visto, de um modelo logit com as variáveis observadas escolhidas nos modelos anteriores. A seguir, a partir de uma média amostral, estima-se o efeito médio do tratamento sem a necessidade de um modelo de regressão linear.

O resultado aplicado ao projeto STAR é:

$$A\hat{T}E_4 = 10,84$$

4.5. Método 5

O método 5 assume a validade das Hipótese 2 e 4. O procedimento consiste na estimação do escore de propensão através de um modelo logit, como já foi feito no método 4, e na regressão da variável resultado *score* na variável tratamento *cltype3* e na variável que estima o escore de propensão.

Criou-se, portanto uma nova variável:

- *p_hat*: variável que estima o escore de propensão, isto é, a probabilidade de seleção de cada indivíduo a partir de um modelo logit (descrito na equação 34)

Variável	Coeficiente	p-valor
Intercepto	1315,49	0
<i>cltype3</i>	11,67	0,01
<i>p_hat</i>	-179,31	4,59E-08
$R^2 = 0,0149$		

Tabela 7 - coeficientes da regressão do resultado *score* no tratamento *cltype3* e na estimativa do escore de propensão *p_hat*

Percebe-se que tanto o escore de propensão quanto o tratamento são estatisticamente significantes para o nível de referência considerado. No entanto, o valor do R^2 para esta regressão é muito baixo, o que sugere que a variável resultado não é bem explicada por este modelo.

A estimativa do ATE é muito próxima dos métodos 2 e 3:

$$\hat{ATE}_5 = 11,67$$

4.6. Método 6

O método 6 relaxa a Hipótese 4, assumindo apenas a validade da Hipótese 2. Para tal, deve-se incluir um termo de interação entre o tratamento e a variável estimativa do escore de propensão *p_hat*. Assim, tal como no método 3, uma nova variável deve ser criada para incluir a interação no modelo de regressão.

- *p_hatnorm_int*: variável *p_hat*, que representa a estimativa do escore de propensão de cada aluno, subtraída de sua média e multiplicada pelo tratamento *cltype3*

O modelo agora é a regressão da variável *score* nas variáveis *cltype3* (tratamento), *p_hat* (estimativa do escore de propensão) e *p_hatnorm_int* (interação).

Variável	Coeficiente	p-valor
Intercepto	1328,35	0
<i>cltype3</i>	11,51	1,96E-05
<i>p_hat</i>	-212,7	2,01E-07
<i>p_hatnorm_int</i>	92,91	1,73E-01
$R^2 = 0,0155$		

Tabela 8 - coeficientes da regressão do resultado *score* no tratamento *cltype3*, na estimativa do escore de propensão *p_hat* e na interação entre o tratamento e a estimativa do escore de propensão

Tal como no método 3, o termo de interação não é estatisticamente significativo a nível de 5%. Além disso, o coeficiente do tratamento é semelhante ao método anterior, sem a variável de interação. Isto sugere que, uma vez assumida a validade da Hipótese 2, assumir adicionalmente a validade da Hipótese 4 é pouco restritivo para o modelo e pouco altera o resultado final da estimativa.

A estimativa do ATE através do método é, portanto:

$$\hat{ATE}_6 = 11,51$$

4.7. Método 7

Dos seis métodos aplicados até aqui, o primeiro supôs a improvável validade da Hipótese 1, que implica na total aleatoriedade da atribuição do tratamento, e todos os outros cinco supuseram a validade da Hipótese 2. Todos eles encontraram estimativas para o ATE com valores semelhantes.

Como visto na seção 2.2, a Hipótese 2 assume que, apesar da não aleatoriedade da atribuição do tratamento, as variáveis observadas contidas no vetor x contêm informação suficiente para explicar a atribuição de tratamento. Desta maneira, a inclusão destas variáveis nos modelos de estimação permitiria isolar o efeito apenas do tratamento e estimar o ATE.

No entanto, ao examinar o banco de dados do projeto STAR, não parece razoável que a Hipótese 2 se verifique. Como visto na seção 3.1, o motivo da quebra da aleatoriedade foram as constantes mudanças de turma dos estudantes participantes do estudo, resultado principalmente de problemas comportamentais, mudanças para outro estado ou de pressões por parte dos pais. Nenhuma das variáveis observadas contidas no banco de dados utilizado na estimação através dos métodos 1 a 6 contém informação o suficiente para explicar estas ações que resultaram na mudança de turma. Na verdade, tais fatores são de difícil quantificação.

Desta maneira, é interessante que haja um método de estimação no qual não seja necessário explicar a atribuição de tratamento através de variáveis observadas. Como visto na seção 2.4, o método de estimação através de variáveis instrumentais não assume explicação do tratamento e, portanto, permite relaxar a Hipótese 2.

Para aplicar o método de variáveis instrumentais, porém, é preciso que haja uma variável, chamada de instrumento (ou z , no desenvolvimento teórico), que atenda às duas condições impostas pela Hipótese 5, isto é, que seja exógena ao modelo de regressão original e que seja correlacionada com a variável tratamento.

Como visto na seção 3.1, a divisão original dos estudantes do projeto STAR entre turmas regulares e turmas reduzidas foi totalmente aleatória. Desta maneira, pode-se intuir que a variável que representa esta divisão (presente no banco com o nome de *cltypek*) é exógena ao modelo de regressão aplicado no método 1. Além disso, por mais que muitos alunos tenham mudado de turma com o passar dos anos (e, assim, quebrado a aleatoriedade do experimento), a maioria não o fez e continuou na turma que lhe foi originalmente atribuída. Desta maneira, existe uma correlação entre a atribuição das turmas no jardim de infância e turma dos alunos na terceira série, que é o tratamento.

De fato, pode-se verificar a existência desta correlação examinando os dados do banco:

$$\text{corr}(\text{cltypek}, \text{cltype3}) = 0,6926$$

É possível então utilizar a variável *cltypek* como instrumento para a variável tratamento *cltype3* e obter uma estimação consistente para o efeito médio do tratamento desejado.

O método a ser aplicado, como visto na seção 2.4, é o MQ2E (Mínimos Quadrados em 2 Estágios). Apesar de o método estimar, como visto, dois modelos de regressão interligados, a maioria dos softwares de análise estatística possui incorporada uma função de MQ2E que, a partir das variáveis observadas e do(s) instrumento(s) escolhido(s), calcula os coeficientes de estimação de maneira direta. No *software* R, utilizado na elaboração deste estudo de caso, não é diferente. Os coeficientes finais estão na Tabela 9 a seguir.

Variável	Coeficiente	p-valor
Intercepto	1229,18	0
<i>cltype3</i>	18,17	8,02E-07
<i>ssex</i>	6,09	0,01
<i>srace</i>	28,76	0,00
<i>hdeg3</i>	5,7	0,23
<i>ses3</i>	-33,69	0,00
$R^2 = 0,1213$		

Tabela 9 - Coeficientes da regressão por variáveis instrumentais do resultado *score* no tratamento *cltype3* e nas demais variáveis utilizando como instrumento a variável *cltypek*

Todos os coeficientes resultantes da regressão por variáveis instrumentais são estatisticamente significantes a 5%. Além disso, percebe-se que o coeficiente da variável tratamento, que é a estimativa do ATE, é diferente do valor encontrado através dos outros 6 métodos.

A estimativa do ATE através de variáveis instrumentais é, portanto:

$$\widehat{ATE}_7 = 18,17$$

4.8. Resumo a análise dos resultados

A Tabela 10 resume os resultados encontrados para a estimativa do ATE através dos diferentes métodos utilizados neste trabalho. Uma coluna foi adicionada com o valor do ATE em percentual do total de pontos que um aluno pode obter na prova.

	Método	$\widehat{ATE}$	$\widehat{ATE}$ (%)
1	Diferença simples das médias dos grupos	10,47	0,88%
2	Regressão direta sem termo de interação	11,68	0,97%
3	Regressão direta com termo de interação	11,55	0,96%
4	Função do Escore de Propensão	10,84	0,90%
5	Regressão com Escore de Propensão sem termo de interação	11,67	0,97%
6	Regressão com Escore de Propensão com termo de interação	11,51	0,96%
7	Variáveis Instrumentais	18,17	1,51%

Tabela 10 - Resumo dos resultados encontrados para as estimativas do ATE

Observando os resultados de estimativa para o ATE resumidos na Tabela 10, pode-se notar que os seis primeiros métodos encontram resultados bastante semelhantes. Todos estes métodos, como já dito,

assumem a validade das Hipóteses 1 ou 2, que, no caso do banco de dados do projeto Tennessee STAR, não se verificam empiricamente.

A última estimativa, através do método de variáveis instrumentais, apresenta um resultado significativamente diferente dos outros. Como foi argumentado na seção 2.4, este método relaxa a Hipótese 2 e depende apenas da existência de uma variável que respeite as duas condições da Hipótese 5: exogeneidade ao modelo de regressão e correlação com o tratamento. Na seção 4.7, viu-se que estas duas condições são satisfeitas pela variável *cltypek*, que define o tipo de turma à qual os estudantes foram aleatoriamente selecionados no início do estudo. Esta variável pode, então, ser usada como instrumento para a variável tratamento do problema.

Desta maneira, o método 7, que produz um resultado diferente de todos os anteriores, é o único que estima de maneira não-viesada e consistente o efeito médio do tratamento e, portanto, deve ser usado para todos os fins práticos nos dados do projeto Tennessee STAR.

Duas questões pertinentes à relevância prática do experimento Tennessee STAR podem ser abordadas à luz dos resultados encontrados.

Primeiro, apesar da significância estatística do valor do ATE ter sido comprovada em todos os métodos de estimação, seu valor percentual pode ser considerado baixo. No caso mais preciso (e com o maior valor estimado para o ATE, o método 7), o aumento médio na nota de um aluno é de apenas 1,51%. Considerando a complexidade, os desafios operacionais e o custo financeiro da implementação da medida proposta, questiona-se se o efeito na redução das turmas dos alunos das primeiras séries primárias, apesar de claramente significativo em termos estatísticos, é relevante. De fato, o experimento foi realizado nos anos 80 e, até o presente momento, a redução no tamanho das turmas não se tornou uma política muito comum em escolas norte-americanas.

Cabe também uma discussão a respeito da variável resultado escolhida para o experimento. Como definido no Capítulo 3, a variável resultado score representa a soma das notas dos alunos participantes em testes padronizados de inglês e matemática. No entanto, pode-ser argumentar que a simples soma de notas em disciplinas distintas como são as duas em questão faz com que se perda algo da objetividade da avaliação das habilidades que as notas individuais das duas provas deseja aferir.

De fato, pode-se ir mais fundo nessa discussão e questionar de maneira ainda mais geral a maneira como se mede a qualidade da educação. Como abordado em Krueger (1997), a escolha de uma variável que represente de maneira adequada os benefícios educacionais obtidos a partir de uma dada política é altamente subjetiva e objeto de intenso debate entre especialistas da área. Certamente, uma única nota em um teste, por mais bem elaborado que seja, não é o suficiente para quantificar a qualidade geral da educação dada a um grupo de crianças. Na verdade, pode-se argumentar que, ao contrário do que o projeto STAR quantifica, turmas maiores trazem benefícios aos alunos no que diz respeito à introdução dos alunos à diversidade humana e à vida em sociedade, fatores que, sem dúvida, são de grande importância no desenvolvimento de uma criança em idade escolar.

Na primeira parte deste trabalho, foi apresentado o desenvolvimento teórico de diversos métodos diferentes de estimação do Efeito Médio do Tratamento (ATE). Cada um dos métodos apresentados tem diferentes níveis de complexidade teórica e apresenta vantagens e desvantagens de ordem operacional.

Cada um dos métodos apresentados assume hipóteses empíricas acerca da base de dados na qual a metodologia será aplicada. Como a violação de alguma das hipóteses necessárias faz com que a estimação final seja viesada, é fundamental que o método escolhido seja coerente com que os dados do problema de modo a que as hipóteses empíricas assumidas sejam respeitadas.

Em um experimento aleatório controlado perfeito, a estimação do ATE é trivial (método 1 da Tabela 2). No entanto, em aplicações em análise de causalidade realizadas *a posteriori*, em experimentos não controlados ou em casos em que a aleatoriedade foi violada (como no estudo de caso apresentado), assumir aleatoriedade perfeita (Hipótese 1) leva a um viés no estimador. Nessas condições, o uso do método 1 é incorreto.

Há casos em que, embora não haja aleatoriedade no tratamento, é possível, através de características adicionais dos indivíduos estudados, explicar a maneira pela qual o tratamento foi atribuído (Hipótese 2). Para estes casos, o ATE pode ser consistentemente estimado através dos métodos 2 a 6, que combinam modelos de regressão linear e pareamento de indivíduos através do escore de propensão. Foi visto também que, uma vez respeitada a Hipótese 2, as Hipóteses 3 e 4, que simplificam os modelos econométricos, podem ser assumidas sem perda significativa.

Finalmente, há o caso mais geral em que não há aleatoriedade e não é possível explicar o tratamento através de outras variáveis observadas. Nesta situação, usa-se o método de variáveis instrumentais para estimar consistentemente o efeito do tratamento. Para que este método seja aplicado, a única condição é a existência de um instrumento adequado, isto é, respeitando as condições impostas pela Hipótese 5.

Na segunda parte deste trabalho, as metodologias 1 a 7 foram aplicadas a um estudo de caso com dados do Projeto Tennessee STAR, realizado em escolas públicas do estado americano do Tennessee, nos anos 1980, com o objetivo de inferir o efeito da redução no tamanho das turmas o aprendizado de alunos dos quatro primeiros anos escolares.

Neste caso, o tratamento cujo efeito foi estimado foi a atribuição de cada aluno a uma turma de tamanho reduzido, isto é, com menos alunos que as turmas regulares das escolas em questão.

Os métodos que assumiram aleatoriedade perfeita (1) ou a possível explicação do tratamento (2 a 6) produziram estimativas semelhantes para o ATE. Vimos, porém, que, embora o experimento tenha sido desenhado para respeitar aleatoriedade total na atribuição do tratamento, diversos fatores de ordem prática tiveram como resultado a quebra desta aleatoriedade planejada inicialmente. Além disso, os fatores causadores desta quebra de aleatoriedade não foram quantificados em qualquer variável observada na população.

Desta maneira, o único método consistente de estimação do efeito do tratamento é o método 7, que se vale da teoria de variáveis instrumentais para relaxar as hipóteses anteriores e permitir quebra de aleatoriedade sem explicação da atribuição do tratamento. De fato, o resultado obtido pela estimação através do método 7 é significativamente diferente (e, portanto, mais correto) dos outros seis.

A partir dos resultados do experimento apresentados na Tabela 10, pode-se questionar a relevância prática da política sugerida como benéfica pelo experimento Tennessee STAR. Embora fique claro que os resultados para a estimação do ATE sejam estatisticamente significantes e positivos, um aumento médio de apenas 1,51% no resultado total dos testes padronizados aplicados aos alunos pode não ser considerado relevante o suficiente para justificar a aplicação de uma política deste tipo em larga escala.

De fato, como discutido na seção 4.8, cabe um questionamento mais profundo a respeito da maneira como se quantifica um conceito tão subjetivo quanto a qualidade da educação dada a um grupo de alunos.

A análise do Projeto Tennessee STAR exemplifica a diversidade de aplicações práticas do conceito de efeito médio do tratamento. Luiz e Struchiner (2002) trabalham este conceito na área de epidemiologia e mostram sua fundamental importância.

Aqui, o foco do conceito de ATE foi mais direcionado para a análise de resultados e planejamento de implementação de políticas públicas, como no caso da política educacional implementada no Tennessee pelo projeto STAR. De fato, Heckman (2005) e Angrist e Pischke (2009) consideram o ATE uma das principais ferramentas de análise científica de causalidade em políticas públicas.

Estas duas áreas são apenas algumas das inúmeras nas quais o conceito de ATE pode ser aplicado em análise da causalidade. Para outros exemplos, além de deduções matemáticas ausentes neste trabalho e discussões mais aprofundadas de muito do que foi aqui abordado, o leitor deve consultar as obras sugeridas no Capítulo 6.

6. Referências Bibliográficas

ANGRIST, J.; PISCHKE, J.S. (2009) *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, Princeton, NJ.

HECKMANN, J.J. (2005) *The Scientific Model of Causality*. In: *Sociological Methodology Vol 35 Issue 1*. Wiley Online Library.

HOLLAND, P.W. (1986) *Statistics and Causal Inference*. In: *Journal of the American Statistical Association*, Vol. 81, No. 396, 945-960.

KRUEGER, A.B. (1997) *Experimental Estimates of Education Production Function*. Natioal Bureau of Economic Research, working paper number 6051.

LUIZ, R.R.; STRUCHINER, C.J. (2002) *Inferência Causal em Epidemiologia: o modelo de respostas potenciais*. Editora Fiocruz, Rio de Janeiro.

ROSENBAUM, P.R.; RUBIN, D.B. (1983) *The central role of the propensity score in observational studies for causal effects*. In: *Biometrika* 70(1), 41-55.

STOCK, J.H.; WATSON, M. W. (2003) *Introduction to Econometrics*. Pearson Education, Inc. Estados Unidos da América.

WOOLDRIDGE, J.M. (2000) *Introductory Econometrics: A Modern Approach*. South-Western College Publishing, Thomson Learning, Estados Unidos da América.

WOOLDRIDGE, J.M. (2010) *Econometric Analysis of Cross-Section and Panel Data*. 2a Edição. Massachusetts Institute of Technology, Estados Unidos da América.

ANEXO – Código em R para o estudo de caso

```
##pré-processamento dos dados do banco original##

data_star = read.csv("Star.csv")

#apenas as variáveis relacionadas à terceira série além da dummy para seleção no jardim são
selecionadas

data1 <- subset(data_star, stark == 'yes' & star3 == 'yes', select =
c(ssex,srace,cltypek,cltype3,schtype3,treadss3,tmathss3,ses3,trace3,hdeg3,totexp3))

data1 <- data1[complete.cases(data1),]

#substituição de valores categóricos por valores numéricos

#tipo de turma no jardim: 0-turma regular, 1-turma reduzida

levels(data1$cltypek)[1] <- "temp"

data1$cltypek<-sapply(data1$cltypek,switch,"temp"=0,"regular + aide class"=0,"regular
class"=0,"small class"=1)

#tipo de turma na terceira série: 0-turma regular, 1-turma reduzida

levels(data1$cltype3)[1] <- "temp"

data1$cltype3<-sapply(data1$cltype3,switch,"temp"=0,"regular + aide class"=0,"regular
class"=0,"small class"=1)

#almoço: 0-nao recebe almoço gratuito, 1-recebe almoço gratuito

levels(data1$ses3)[1] <- "temp"

data1$ses3<-sapply(data1$ses3,switch,"temp"=0,"free lunch"=1,"non-free lunch"=0)

#etnia do aluno: 0-aluno não é branco, 1-aluno é branco (quase todos os não-brancos são negros)

levels(data1$srace)[1] <- "temp"

data1$srace<-
sapply(data1$srace,switch,"temp"=0,"am.indian"=0,"asian"=0,"black"=0,"hispanic"=0,"other"=0,"whit
e"=1)

#etnia do professor: 0-professor não é branco, 1-professor é branco (quase todos os não-brancos são
negros)

levels(data1$trace3)[1] <- "temp"

data1$trace3<-sapply(data1$trace3,switch,"temp"=0,"asian"=0,"black"=0,"white"=1)

#sexo do aluno: 0-masculino, 1-feminino

levels(data1$trace3)[1] <- "temp"

data1$ssex<-sapply(data1$ssex,switch,"temp"=0,"female"=1,"male"=0)
```

#qualificação do professor: 0-até graduação, 1-pós-graduação (mestrado ou especialização)

```
levels(data1$hdege3)[1] <- "temp"
```

```
data1$hdeg3<-sapply(data1$hdeg3,switch,"temp"=0,"bachelors (bs/ba)"=0,"masters  
(ms/ma/med)"=1,"specialist (ed.s)"=1)
```

#tipo de escola: 0-urbana, 1-rural

```
levels(data1$schtype3)[1] <- "temp"
```

```
data1$schtype3<-sapply(data1$schtype3,switch,"temp"=0,"inner-city"=0,"rural school"=1,"suburban  
school"=0,"urban school"=1)
```

436 (14.5%) das crianças mudaram o tipo de turma entre o jardim e a 3a série

#convertendo a variável experiência para tipo numérico

```
data1$totexp3 = as.numeric(data1$totexp3)
```

#criação da variável "score" que soma as pontuações nos testes de inglês e matemática

```
data1["score"]<-NA
```

```
data1$score <- data1$treadss + data1$tmathss
```

#Imprime o novo banco de dados

```
write.csv(data1,file = "Star2.csv")
```

##Estimação por diferença de médias##

```
data = read.csv("Star2.csv")
```

```
data0 <- subset(data,data$cltype3 == 0)
```

```
data1 <- subset(data,data$cltype3 == 1)
```

```
mean(data1$score) - mean(data0$score)
```

```
#ATE_hat = 10.46954
```

##Estimação por regressão sem termo de interação##

```
data = read.csv("Star2.csv")
```

```
reg1 <- lm(data$score ~ data$cltype3 + data$ssex + data$trace3 + data$srace + data$totexp3 +  
data$ses3 + data$hdeg3 + data$schtype3)
```

```
coef(summary(reg1))
```

#A experiência do professor não é significativa, refazemos a regressão sem ela

```
reg2 <- lm(data$score ~ data$cltype3 + data$ssex + data$srace + data$hdeg3 + data$ses3)
```

```
coef(summary(reg2))
```

#Estimativa do ATE é o coeficiente que multiplica a variável tratamento (no caso cltype3)

#ATE_hat = 11.677269

##Estimação por regressão com termo de interação##

data = read.csv("Star2.csv")

#Normalizando para os termos de interação

data["ssexnorm"]<-NA

data\$ssexnorm <- data\$ssex - mean(data\$ssex)

data["sracenorm"]<-NA

data\$sracenorm <- data\$srace - mean(data\$srace)

data["hdeg3norm"]<-NA

data\$hdeg3norm <- data\$hdeg3 - mean(data\$hdeg3)

data["ses3norm"]<-NA

data\$ses3norm <- data\$ses3 - mean(data\$ses3)

#Criando as variáveis de interação

data["ssexnorm_int"]<-NA

data\$ssexnorm_int <- data\$ssexnorm*data\$cltype3

data["sracenorm_int"]<-NA

data\$sracenorm_int <- data\$sracenorm*data\$cltype3

data["hdeg3norm_int"]<-NA

data\$hdeg3norm_int <- data\$hdeg3norm*data\$cltype3

data["ses3norm_int"]<-NA

data\$ses3norm_int <- data\$ses3norm*data\$cltype3

reg <- lm(data\$score ~ data\$cltype3 + data\$ssex + data\$srace + data\$hdeg3 + data\$ses3 + data\$ssexnorm_int + data\$sracenorm_int + data\$hdeg3norm_int + data\$ses3norm_int)

coef(summary(reg))

#Nenhuma das variáveis de interação é estatisticamente significativa

#Estimativa do ATE é o coeficiente que multiplica a variável tratamento (no caso cltype3)

#ATE_hat = 11.554083

##Estimação por função do propensity score##

data = read.csv("Star2.csv")

#Estimando através de um modelo logit a probabilidade de seleção de cada indivíduo

```
p_hat <- glm(data$cltype3 ~ data$ssex + data$srace + data$hdeg3 + data$ses3, family =  
binomial(link = logit))
```

#Coeficientes para a estimação

```
b0 = p_hat$coeff[[1]]
```

```
b1 = p_hat$coeff[[2]]
```

```
b2 = p_hat$coeff[[3]]
```

```
b3 = p_hat$coeff[[4]]
```

```
b4 = p_hat$coeff[[5]]
```

#Valor estimado da probabilidade

```
data["p_hat"]<-NA
```

```
data$p_hat <- exp(b0 + b1*data$ssex + b2*data$srace + b3*data$hdeg3 + b4*data$ses3)/(1 +  
exp(b0 + b1*data$ssex + b2*data$srace + b3*data$hdeg3 + b4*data$ses3))
```

#Criação da variável cuja média é a estimativa do ATE

```
data["ATEind"]<-NA
```

```
data$ATEind <- ((data$cltype3 - data$p_hat)*data$score)/(data$p_hat*(1-data$p_hat))
```

#Estimando o ATE a partir da média da variável criada

```
ATE_hat = mean(data$ATEind)
```

```
#ATE_hat = 10.84514
```

##Estimação através da regressão incluindo o propensity score sem termo de interação##

```
data = read.csv("Star2.csv")
```

#Estimando através de um modelo logit a probabilidade de seleção de cada indivíduo

```
p_hat <- glm(data$cltype3 ~ data$ssex + data$srace + data$hdeg3 + data$ses3, family =  
binomial(link = logit))
```

#Coeficientes para a estimação

```
b0 = p_hat$coeff[[1]]
```

```
b1 = p_hat$coeff[[2]]
```

```
b2 = p_hat$coeff[[3]]
```

```
b3 = p_hat$coeff[[4]]
```

```
b4 = p_hat$coeff[[5]]
```

#Valor estimado da probabilidade

```
data["p_hat"]<-NA
```

```
data$p_hat <- exp(b0 + b1*data$ssex + b2*data$srace + b3*data$hdeg3 + b4*data$ses3)/(1 +  
exp(b0 + b1*data$ssex + b2*data$srace + b3*data$hdeg3 + b4*data$ses3))
```

```
#regressão da variável resposta em w e p_hat
```

```
reg <- lm(data$score ~ data$cltype3 + data$p_hat)
```

```
coef(summary(reg))
```

```
#Estimativa do ATE é o coeficiente da variável que multiplica o tratamento cltype3
```

```
#ATE_hat = 11.67268
```

```
##Estimação através da regressão incluindo o propensity score com termo de interação##
```

```
data = read.csv("Star2.csv")
```

```
#Estimando através de um modelo logit a probabilidade de seleção de cada indivíduo
```

```
p_hat <- glm(data$cltype3 ~ data$ssex + data$srace + data$hdeg3 + data$ses3, family =  
binomial(link = logit))
```

```
#Coeficientes para a estimação
```

```
b0 = p_hat$coeff[[1]]
```

```
b1 = p_hat$coeff[[2]]
```

```
b2 = p_hat$coeff[[3]]
```

```
b3 = p_hat$coeff[[4]]
```

```
b4 = p_hat$coeff[[5]]
```

```
#Valor estimado da probabilidade
```

```
data["p_hat"]<-NA
```

```
data$p_hat <- exp(b0 + b1*data$ssex + b2*data$srace + b3*data$hdeg3 + b4*data$ses3)/(1 +  
exp(b0 + b1*data$ssex + b2*data$srace + b3*data$hdeg3 + b4*data$ses3))
```

```
#Normalizando a variável p_hat para criar a interação com w
```

```
data["p_hatnorm"]<-NA
```

```
data$p_hatnorm <- data$p_hat - mean(data$p_hat)
```

```
#Criação da variável interação
```

```
data["p_hatnorm_int"]<-NA
```

```
data$p_hatnorm_int <- data$p_hatnorm*data$cltype3
```

```
#Regressão da variável resposta em w, p_hat e na interação dos dois
```

```
reg <- lm(data$score ~ data$cltype3 + data$p_hat + data$p_hatnorm_int)
```

```
coef(summary(reg))
```

```
#Coeficiente da interação não é significativo
```

```
#ATE = 11.51499
```

```
##Estimação por variáveis instrumentais##
```

```
data = read.csv("Star2.csv")
```

```
reg1 <- lm(data$score ~ data$cltype3 + data$ssex + data$srace + data$hdeg3 + data$ses3)
```

```
coef(summary(reg1))
```

```
#estimador é viesado
```

```
install.packages("sem")
```

```
library("sem")
```

```
regiv <- tsls(score ~ cltype3 + ssex + srace + hdeg3 + ses3, ~ cltypek + ssex + srace + hdeg3 +  
ses3, data=data)
```

```
coef(summary(regiv))
```

```
#Estimativa do ATE é o coeficiente que multiplica a variável tratamento (cltype3)
```

```
#ATE_hat = 18.171076
```