3
Estimação por Imputação e Reponderação

Neste capítulo serão discutidos dois dos métodos que têm recebido maior atenção da literatura de estimação de efeito de tratamento sob ignorabilidade[1].

O primeiro, consiste na imputação do valor estimado da função de regressão no lugar dos valores potenciais, viabilizando a estimação do Efeito Médio de Tratamento por uma média de diferenças. O outro envolve a comparação dos valores médios da variável de interesse entre os grupos, sendo estes devidamente reponderados de forma a representar a população.

Apesar das inspirações distintas, as duas abordagens levam a estimadores que podem ser analisados através de uma teoria geral como a desenvolvida em Newey (1994). Esta teoria garante que ambas possuem boas propriedades assintóticas, sob hipóteses adicionais de suavidade das funções envolvidas e regularidade do modelo probabilístico subjacente.

3.1
Regressão/Imputação

Uma abordagem para a estimação do Efeito Médio de Tratamento é motivada pela equação (2-3), a qual pode ser interpretada como uma condição de momento associada a funções que podem ser estimadas com os dados disponíveis. Isto sugere que estimemos β por:

$$\hat{\beta}_{imp} = \frac{1}{N}\sum_{i=1}^{N} \hat{m}_1(X_i) - \hat{m}_0(X_i), \tag{3-1}$$

onde $\hat{m}_t(.)$ é uma estimativa para $m_t(.)$. O procedimento envolve, portanto, dois passos: no primeiro, estimar as funções de regressão; no segundo, integrar a diferença entre as estimativas sobre a distribuição empírica de X_i. Intuitivamente, este estimador substitui a diferença entre valores potenciais da variável de interesse, que requer o uso de dados indisponíveis, pela diferença entre valores das funções de regressão, que podem ser estimados e imputados.

[1]Na resenha de Imbens e Wooldridge (2009), os principais métodos são classificados em três grupos: regressão, métodos baseados no *propensity score*, e *matching*. Neste trabalho são abordadas a técnica de regressão e a de reponderação, que se enquadra no segundo grupo. Os métodos de *matching*, ineficientes em geral, não serão discutidos aqui por não possuirem ligação próxima com a proposta deste estudo.

O método de imputação foi implementado por Rubin (1977), num contexto em que $m_1(.)$ e $m_0(.)$ seguem um modelo paramétrico previamente conhecido. Mais recentemente, foi estudada a implementação no caso de interesse para este trabalho, em que as formas funcionais das funções de regressão são desconhecidas. Heckman, Ichimura e Todd (1997), e Heckman, Ichimura, Smith e Todd (1998) empregaram métodos de *kernel* para a primeira etapa.

Outra possibilidade é obter $\hat{m}_t$ por um método de *sieves* (Grenander (1981), Geman e Hwang (1982)), que consiste em estimar modelos paramétricos, porém sucessivamente mais abrangentes quanto maior a amostra. Um exemplo é a estimação por *sieve*/mínimos quadrados, em que se considera uma seqüência crescente de bases $B_K = \{q_k(X), k = 1, 2, ..., K\}$ de funções de X, e, para cada tamanho de amostra n, $\hat{m}_t(X_i) = q^{k(n)}(X_i)'\hat{\gamma}_{k(n),t}$, onde $\hat{\gamma}_{k,t}$ são os coeficientes da projeção ortogonal de Y sobre $q_1(X), ..., q_{k(n)}(X)$ na sub-amostra $\{(Y_i, X_i, T_i); T_i = t\}$. Ou seja, introduzindo a notação $(X_i^1, Y_i^1)_{i=1}^{N_1}$, para a subamostra das unidades tratadas, e $(X_i^0, Y_i^0)_{i=1}^{N_0}$, das de controle,

$$
\begin{aligned}
q^k(X) &= (q_1(X), ..., q_k(X)) \\
\hat{\gamma}_{k,t} &= \left(Q^{k,t\prime} Q^{k,t}\right)^{-1} Q^{k,t\prime} Y^t \\
Q^{k,t} &= (q^k(X_1^t)', ..., q^k(X_{n_t}^t)').
\end{aligned}
$$

Como requisito mínimo para a convergência de $\hat{m}_t$, $m_t(.)$ deve ser aproximada arbitrariamente bem no espaço gerado por B_K, para K suficientemente grande, e a dimensão $k(n)$ do espaço de projeção deve crescer arbitrariamente com, porém menos que proporcionalmente a, o tamanho da amostra.

Esta técnica foi adotada por Chen, Hong e Tamer (2005), num problema de estimação similar ao nosso[2]. Imbens, Newey e Ridder (2005) estudaram a versão do mesmo estimador na qual as B_K são bases de polinômios, no contexto de estimação de Efeito Médio de Tratamento sob Ignorabilidade.

3.2

[2]Chen, Hong e Tamer (2005) consideram o problema de estimar a solução de uma condição de momento envolvendo variáveis latentes, X^*, observadas apenas numa amostra auxiliar. A hipótese de identificação, de que a distribuição de X^* condicional às variáveis sempre observadas X independe de a observação estar na amostra auxiliar, é análoga à Ignorabilidade, com $X^* = Y(t)$ e as unidades que recebem $T = t$ correspondendo à amostra auxiliar na estimação de $E[Y(t)]$.

Reponderação

Um grupo de métodos alternativos à Regressão/Imputação se desenvolveu ligado ao trabalho de Rosenbaum e Rubin (1983, 1984), que mostraram que o *propensity score* contém a informação necessária sobre a diferença entre os grupos de tratamento e controle. Formalmente, Rosenbaum e Rubin demonstram que, sob Ignorabilidade:

$$T \perp (Y(0), Y(1)) \mid p(X), \tag{3-2}$$

e que, portanto, não há viés em se comparar unidades com o mesmo *propensity score*. Esta observação motivou a busca por estimadores baseados nessa função, evitando métodos de imputação que, por envolverem uma estimação de dimensão alta na primeira etapa, freqüentemente apresentavam dificuldades computacionais e mau desempenho em amostras pequenas.

De particular interesse para este trabalho é o método que utiliza o *propensity score* numa reponderação das unidades, como na segunda fórmula de identificação, (2-5). A reponderação pode ser interpretada como uma forma de equilibrar a distribuição de X entre os grupos na amostra.

Procedimentos de reponderação foram inicialmente empregados em contextos nos quais o *propensity score* era conhecido. Horwitz e Tompson (1952), freqüentemente apontados como precursores da técnica, utilizaram a ponderação pelo inverso da probalidade de seleção em amostras estratificadas.

No caso da estimação do Efeito Médio de Tratamento com dados observacionais, o *propensity score* geralmente é desconhecido, e portanto deve ser estimado. Por analogia à equação (2-5), um estimador de Reponderação tem a forma:

$$\hat{\beta}_{rep} = \frac{1}{N} \sum_{i=1}^{N} \frac{T_i Y_i}{\hat{p}(X_i)} - \frac{(1 - T_i) Y_i}{1 - \hat{p}(X_i)}, \tag{3-3}$$

ou

$$\tilde{\beta}_{rep} = \frac{\sum_{i=1}^{N} T_i Y_i / \hat{p}(X_i)}{\sum_{i=1}^{N} T_i / \hat{p}(X_i)} - \frac{\sum_{i=1}^{N} (1 - T_i) Y_i / (1 - \hat{p}(X_i))}{\sum_{i=1}^{N} (1 - T_i) / (1 - \hat{p}(X_i))}, \tag{3-4}$$

onde $\hat{p}(.)$ é uma estimativa do *propensity score*. A diferença entre os dois estimadores é que $\tilde{\beta}_{rep}$ impõe que a soma dos pesos valha um. Se uma normalização dos pesos for incluída no primeiro passo, pela forma como $\hat{p}(.)$ é estimado, temos $\hat{\beta}_{rep} = \tilde{\beta}_{rep}$.

Como o *propensity score* corresponde a uma média condicional, pode ser estimado por métodos de *sieve*/mínimos quadrados, da mesma forma que as funções de regressão no estimador de Imputação discutido acima. As propriedades de um estimador deste tipo foram estudadas por Hirano, Imbens e Ridder (2000), com bases polinomiais e, mais tarde, por Chen, Hong e Tarozzi (2008), de forma mais geral.

No entanto, o emprego do método de mínimos quadrados permite que os valores estimados $\hat{p}(X_i)$ não pertençam ao intervalo unitário, e portanto se produzam pesos negativos[3]. Devido a esta desvantagem, Hirano, Imbens e Ridder (2003) propõem estimar $p(.)$ por um método *sieve* alternativo. No lugar da projeção por mínimos quadrados, utilizam a estimação por máxima verossimilhança da regressão logística de T nas funções da base $B_{k(n)}$, ou seja, $\hat{p}(X_i) = L(q^{k(n)}(X_i)'\gamma_{k(n)})$, onde $L(.)$ é a função logística e

$$\hat{\gamma_k} = \arg\max_{\gamma \in \mathbb{R}^k} \sum_{i=1}^{N} T_i \log(L(q^k(X_i)'\gamma)) + (1 - T_i)(1 - \log(L(q^k(X_i)'\gamma)).$$

Um fato interessante, observado em diversos trabalhos sobre o assunto (ver, por exemplo, Hahn (1998), Hirano, Imbens e Ridder (2003)), é que a ponderação utilizando o verdadeiro $p(.)$ não permite atingir o limite de eficiência, enquanto isto é possível empregando valores estimados. Por outro lado, a necessidade de obter estas estimativas torna a Reponderação praticamente tão complexa quanto a Imputação, ao contrário do sugerido pela motivação inicial para o estudo de métodos baseados no *propensity score*.

3.3 Propriedades Assintóticas

Ambas as abordagens apresentadas nas seções anteriores envolvem procedimentos em duas etapas: uma estimação não-paramétrica e uma resolução de condição de momento utilizando a estimativa obtida preliminarmente. Estimadores deste tipo são objeto de particular interesse entre os métodos semi-paramétricos, e diversos estudos propõem estruturas gerais para analisá-los, como Bickel et al. (1993), Newey (1994), Newey e McFadden (1994) e Chen, Linton e Van Keilegom (2003).

Na terminologia dessa literatura, a função estimada no primeiro passo é denominada parâmetro de ruído (*nuisance parameter*), em contraste com o parâmetro de interesse. Uma questão fundamental é a forma como a estimação deste é afetada pelo desconhecimento do primeiro, tomando como referência o estimador de método dos momentos obtido caso aquela função fosse conhecida.

Por resultados bem conhecidos da teoria sobre o método dos momentos, a consistência dos estimadores é implicada pela convergência da condição de momento amostral, uniformemente com respeito ao parâmetro de interesse,

[3]Uma crítica semelhante se aplica ao estimador do primeiro passo da Imputação quando Y é variável limitada.

para seu valor populacional. Isto é, no caso da Imputação, $\hat{\beta} \xrightarrow{p} \beta$ se

$$\sup_{\beta}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\hat{m_1}(X_i) - \hat{m_0}(X_i)\right) - \beta - \left(E\left[m_1(X) - m_0(X)\right] - \beta\right)\right]$$
$$= \frac{1}{N}\sum_{i=1}^{N}\left(\hat{m_1}(X_i) - \hat{m_0}(X_i)\right) - E\left[m_1(X) - m_0(X)\right] \xrightarrow{p} 0 \quad (3\text{-}5)$$

e, no caso da Reponderação, se

$$\frac{1}{N}\sum_{i=1}^{N}\frac{T_iY_i}{\hat{p}(X_i)} - \frac{(1-T_i)Y_i}{1-\hat{p}(X_i)} - E\left[\frac{TY}{p(X)} - \frac{(1-T)Y}{1-p(X)}\right] \xrightarrow{p} 0. \quad (3\text{-}6)$$

Uma forma de demonstrar (3-5) e (3-6) é observar que as parcelas individuais $\hat{m_1}(X_i) - \hat{m_0}(X_i)$ e $\frac{T_iY_i}{\hat{p}(X_i)} - \frac{(1-T_i)Y_i}{1-\hat{p}(X_i)}$ são contínuas com respeito a $(\hat{m_0}(.), \hat{m_1}(.))$ e $\hat{p}(.)$, se considerarmos para estes últimos a norma do supremo. Portanto, sob hipóteses que garantam a convergência, nessa norma, dos parâmetros de ruído a seus valores verdadeiros, um argumento de Lei Uniforme dos Grandes Números permite verificar esses limites. Newey (1994, Lema 5.2) sugere esta estratégia, que é utilizada por Hirano, Imbens e Ridder (2000) e Chen, Hong e Tamer (2005), respectivamente para Reponderação e Imputação.

O estudo das propriedades relacionadas à $\sqrt{N}$-consistência é facilitado pela observação de Newey (1994), de que, quando o modelo semiparamétrico é suficientemente amplo, qualquer estimador regular assintoticamente linear (RAL) é eficiente. Este resultado é conseqüência de outro (Newey, 1990) que diz que a função influência de um estimador RAL, ψ, satisfaz, no lugar de d, a equação de diferenciabilidade por caminhos (2-6). Logo, usando a notação da seção 2.5, para qualquer submodelo paramétrico regular,

$$E[(\psi - \delta)S'_\theta] = \frac{\partial\beta(\theta_0)}{\partial\theta} - \frac{\partial\beta(\theta_0)}{\partial\theta} = 0, \quad (3\text{-}7)$$

i.e., a diferença $\psi - \delta$ é ortogonal a qualquer escore. A condição de abrangência do modelo semiparamétrico é que o espaço tangente contenha todas as funções de média zero dos dados. Quando válida, temos que $\psi - \delta$ é simultaneamente pertencente e ortogonal ao espaço tangente; portanto, deve ser idênticamente nula.

De fato, o resultado de Scharfstein, Robins e Rotnitzky (1999) apresentado na seção 2.4, conforme os próprios autores apontam[4], mostra que o modelo em que apenas se impõe Ignorabilidade é amplo como requerido por Newey. Logo, independente da abordagem (se por Reponderação ou Imputação), sob

[4]Op. cit., p. 1118.

condições de regularidade, um estimador assintoticamente linear de β terá a função influência eficiente. Em vista disso, duas formas de escrever esta função são interessantes:

$$\begin{aligned}\psi^* &= m_1(X) - m_0(X) - \beta + \frac{T}{p(X)}(Y - m_1(X)) - \frac{1-T}{1-p(X)}(Y - m_0(X)) \\ &= \frac{T}{p(X)}Y - \frac{1-T}{1-p(X)}Y - \beta - (T - p(X))(\frac{m_1(X)}{p(X)} + \frac{m_0(X)}{1-p(X)}).\end{aligned}$$

Em cada uma das formas, o primeiro termo é a função influência dos estimadores de Imputação e Reponderação que seriam obtidos usando-se os valores verdadeiros das funções de regressão e *propensity score*. O segundo termo pode ser visto como uma correção para a estimação dessas funções. Para o caso da estimação por reponderação, o termo de correção é negativamente correlacionado com o primeiro, daí o ganho de eficiência em se estimar o *propensity score*, mesmo quando se conhece esta função.

As propriedades sobre a distribuição assintótica dependem, portanto, da convergência dos 'resíduos':

$$\frac{1}{N}\sum_{i=1}^{N} \hat{m}_1(X_i) - m_1(X_i) - (\hat{m}_0(X_i) - m_0(X_i)),$$

da estimação de $m_t(.)$ na Regressão, e

$$\frac{1}{N}\sum_{i=1}^{N} \frac{T_iY_i}{\hat{p}(X_i)} - \frac{(1-T_i)Y_i}{1-\hat{p}(X_i)} - \left(\frac{T_iY_i}{p(X_i)} - \frac{(1-T_i)Y_i}{1-p(X_i)}\right),$$

da estimação de $p(.)$ na Reponderação, para os respectivos termos de correção a uma taxa mais rápida que $\frac{1}{\sqrt{N}}$. Esta condição impõe restrições adicionais sobre a taxa à qual os estimadores dos parâmetros de ruído devem convergir. No caso da estimação por Reponderação, um desdobramento da ineficiência do uso do *propensity score* verdadeiro é que a convergência da estimativa deste deve, por outro lado, ser lenta o suficiente para deixar o 'resto' de ordem $\frac{1}{\sqrt{N}}$ mostrado acima. Isto torna não-trivial a incorporação de informações adicionais (como restrições de forma funcional ou de suavidade de $p(.)$) que possam auxiliar o primeiro passo desta abordagem[5].

A velocidade de convergência que um método não paramétrico de estimação de uma função pode atingir depende da suavidade do estimando. Para *sieve*/mínimos quadrados, por exemplo, Newey (1997) mostra que as taxas máximas de convergência uniforme e em média quadrática são funções

[5]Chen, Hong e Tarozzi (2008) mostram uma forma de fazê-lo

decrescentes da dimensão do espaço dos regressores e crescentes do número de derivadas contínuas da função estimada. Assim, quanto maior o conjunto de variáveis auxiliares, maior a suavidade a ser exigida das funções desconhecidas.

Para seu estimador de Reponderação, Hirano, Imbens e Ridder (2003) mostram a normalidade e eficiência assintótica supondo que o *propensity score* é $7r$ vezes diferenciável, onde r é a dimensão do espaço das variáveis auxiliares. Chen, Hong e Tamer (2005) demonstram essas propriedades para o estimador de Imputação, desde que $m_t(.)$ pertençam a determinados espaços de Hölder[6] e $\frac{p(.)}{1-p(.)}$ seja suave o suficiente para admitir determinadas aproximações nos espaços *sieve*.

[6]Uma função $f(.)$ pertence ao espaço de Hölder (γ, p) se tem derivadas até a ordem $\lfloor \gamma \rfloor$ (maior inteiro menor ou igual a γ, sua parte inteira), sendo as derivadas de maior ordem funções Hölder-contínuas de expoente $\alpha = (\gamma - \lfloor \gamma \rfloor)$, i.e., com $\frac{\|f(x)-f(y)\|}{\|x-y\|^{\alpha}}$ limitada.