1
Introdução

A identificação do efeito causal de um tratamento ou programa sobre uma variável de interesse é um dos temas principais da literatura econométrica, sendo parte essencial da avaliação de políticas públicas como intervenções ativas no mercado de trabalho (ver, por exemplo, os trabalhos de Heckman, Ichimura e Todd (1997), e Heckman et al. (1998)). A preocupação central no estudo deste problema é a relação entre os componentes não observados que determinam os resultados e aqueles que afetam a participação no programa. Assim, na impossibilidade de realização de experimentos aleatórios, deve haver auto-seleção para o tratamento, que em geral segue padrões desconhecidos pelo econometrista, ocasionando o fenômeno conhecido como viés de seleção, devido ao efeito de tratamento individual ser diferente entre as unidades afetadas e as não afetadas.

Nestas condições, alguma suposição adicional deve ser feita para identificar um parâmetro de interesse. A hipótese de Ignorabilidade (Rosenbaum e Rubin, 1983) afirma que toda a informação relevante sobre a heterogeneidade pode ser captada por variáveis auxiliares observadas para todas as unidades. Em outras palavras, não há viés de seleção sistemático quando comparam-se indivíduos semelhantes quanto a determinadas características. Considerada isoladamente, esta hipótese define um modelo semiparamétrico para a população, ou seja, impõe que a distribuição de probabilidade que a descreve pertença a uma dada classe a qual, embora seja um subconjunto próprio do universo de todas as medidas de probabilidade, é ampla a ponto de não poder ser indexada por um parâmetro de dimensão finita. Quando há interesse na inferência do efeito de tratamento sobre toda a população, com dados não experimentais, este é o principal modelo considerado na literatura que permite identificação exata de um parâmetro.

A teoria sobre estimação semiparamétrica do Efeito de Tratamento sob Ignorabilidade em grandes amostras encontra-se num estágio de desenvolvimento avançado. Uma variedade de métodos bastante distintos entre si foi profundamente estudada, e, para cada um, foram desenvolvidas condições sob as quais a estimação assintoticamente eficiente é assegurada. Dois dos mais

importantes métodos são a Imputação (ou Regressão) e a Reponderação. No primeiro, os dados de cada grupo são usados para estimar a relação entre os valores potenciais e as variáveis auxiliares, a chamada função de regressão; em seguida estas estimativas são usadas como substitutas do valor potencial não observado no outro grupo. O segundo método envolve estimar a relação entre as variáveis auxiliares e a seleção para o tratamento, dada pela probabilidade desta condicionada àquelas, conhecida como *propensity score*. Esta informação descreve a representação relativa dos grupos para cada valor das características auxiliares, permitindo uma reponderação que torna a amostra representativa de uma população na qual a participação foi atribuída aleatoriamente.

Uma peculiaridade da teoria assintótica para alguns problemas semiparamétricos, inclusive este, é que os procedimentos são equivalentes entre si, desde que condições de regularidade relativamente fracas sejam satisfeitas (Newey, 1994). Questões práticas, entretanto, motivam o interesse no desempenho da inferência em amostras pequenas. A fragilidade da relação entre a teoria, predominantemente assintótica, e as propriedades em amostras finitas motivou uma série de estudos empregando simulações. Estes conseguem apontar a importância da forma de implementação dos métodos, além dos méritos relativos de cada um. Em particular, uma questão apresentada recentemente é o potencial benefício em se integrar diferentes técnicas em procedimentos que compartilhem o bom desempenho de cada uma delas em certas circunstâncias.

Este trabalho visa a um melhor entendimento das possibilidades oferecidas pela combinação das técnicas de Imputação e Reponderação. Para atingir este objetivo, foram realizados dois tipos de esforço. Por um lado, razões teóricas para a superioridade da abordagem mista, ligadas à literatura de estimação duplamente robusta (Robins e Rotnitzky, 1995, Robins, Rotnitzky e Zhao , 1995), são discutidas. É importante ressaltar, entretanto, que a motivação principal para a combinação em nosso contexto se distingue daquela considerada em grande parte dessa produção científica. De fato, a inferência duplamente robusta tem sido aplicada, na maioria dos estudos, para combinar estimadores paramétricos de Imputação e Reponderação de modo a assegurar consistência, desde que a especificação usada em ao menos um deles descreva corretamente o modelo populacional. No presente estudo, pelo contrário, partimos do modelo semiparamétrico onde apenas se supõe Ignorabilidade, e cada abordagem, isoladamente, produz estimadores consistentes sob condições gerais. Procuramos então melhorar o desempenho em amostras finitas pela combinação de Imputação e Reponderação semiparamétricas, possibilidade que, embora apoiada pelo trabalho de Robins e Ritov (1997), permanece pouco explorada na literatura. Com base nessa discussão, propomos

dois procedimentos Duplamente Robustos. O primeiro é uma generalização direta do estimador de Scharfstein, Robins e Rotnitsky (1999), na qual usamos estimativas preliminares do tipo *sieve* no lugar das paramétricas. Este método, que até recentemente não havia sido considerado de forma explícita na literatura, coincide com o estudado por Cattaneo (2007). O outro consiste num procedimento semiparamétrico de Imputação no qual estimativas das funções de regressão são obtidas por mínimos quadrados ponderados pelo inverso do *propensity score* estimado (ou de seu complemento). Este estimador foi implementado por Hirano e Imbens (2001), mas carece de um estudo detalhado de suas propriedades.

Pelo outro lado, de maneira complementar à análise teórica, são realizadas simulações de Monte Carlo comparando diferentes implementações de métodos de Imputação, Reponderação, e combinações destes na forma dos estimadores Duplamente Robustos propostos. Como forma de avaliar a relevância destes últimos, bem como testar previsões e sugestões da teoria, este exercício é reproduzido sob diversos modelos populacionais. As especificações consideradas se distinguem quanto à forma funcional do *propensity score* e das funções de regressão, a heterocedasticidade dos valores potenciais e a dimensão do conjunto de variáveis auxiliares. Quanto ao primeiro atributo, buscamos variar a suavidade do modelo de uma maneira sistemática, apoiada em um conceito 'função suave' relevante para a teoria, evitando assim manipulações *ad hoc*, freqüentes em estudos de simulação. Da mesma forma, a heterocedasticidade foi introduzida de modo a gerar cenários extremamente favoráveis e desfavoráveis para o estimador Duplamente Robusto baseado em Imputação, conforme se alinha ou se opõe a ponderação empregada neste método àquela que otimiza a estimação das funções de regressão.

Os resultados mostram que a combinação de Imputação e Reponderação em procedimentos duplamente robustos permite a redução do erro quadrático médio mesmo para modelos desfavoráveis. Adicionalmente, corroborando a análise teórica, o ganho de eficiência possibilitado pelos estimadores Duplamente Robustos é maior em modelos menos suaves e quando as variáveis auxiliares são multidimensionais. Mudanças na heterocedasticidade, por sua vez, não produzem efeito sobre a vantagem da integração de técnicas, diferentemente do esperado. Por fim, a combinação da estimação de funções de regressão com a ponderação por funções do *propensity score* exato mostrou-se um meio efetivo, na maior parte dos modelos, de aproveitamento do conhecimento prévio desta função.