Introdução à Inferência Causal

Rafael Izbicki*

4 de fevereiro de 2022

Por favor, envie correções e sugestões para rafaelizbicki@gmail.com.

Agradecimentos. Agradeço a Carlos Cinelli, Marcel Ribeiro-Dantas e Victor Reis pelas sugestões feitas.

Figura 1: Estimar o efeito causal de uma variável em outra não é trivial e requer muitas suposições. Uma análise ingênua pode gerar grandes erros.

*http://www.rizbicki.ufscar.br/

Conteúdo

1 Introdução

Em problemas de predição (aprendizado supervisionado), desejamos prever uma variável resposta Y após observar covariáveis $\mathbf{x}$. Para isso, estimamos quantidades associadas à distribuição de probabilidade $Y|\mathbf{X} = \mathbf{x}$, como, por exemplo, a média condicional $r(\mathbf{x}) := \mathbb{E}[Y|\mathbf{x}]$ (Izbicki and dos Santos, 2020). Em inferência causal, nosso objetivo é outro: queremos prever o valor de Y quando *fazemos uma intervenção* de modo a forçar $\mathbf{X} = \mathbf{x}$. Esses objetivos são diferentes, pois a intervenção pode alterar a medida de probabilidade associada ao problema. Vejamos um exemplo que ilustra essa diferença.

Exemplo 1 (Paradoxo de Simpson). Este exemplo é inspirado em um análise de Jeffrey Morris[1]. Seja $H = \mathbb{I}$(hospitalizado) a variável que indica se um indivíduo é hospitalizado por COVID durante o período do estudo, $V = \mathbb{I}$(vacinado) se ele tomou vacina e $I = \mathbb{I}$(idoso) sua faixa etária. Vamos assumir que idosos se vacinam mais que jovens: $\mathbb{P}(V = 1|I = 0) = 0.5$, $\mathbb{P}(V = 1|I = 1) = 0.9$, e que 50% da população é idosa, de modo que o Teorema de Bayes implica que $\mathbb{P}(I = 0|V = 1) \approx 0.357$ e $\mathbb{P}(I = 0|V = 0) \approx 0.833$. Além disso, vamos assumir que a vacina protege tanto jovens quanto idosos, mas que idosos têm maior probabilidade de hospitalização, com ou sem vacina:

$$\begin{aligned}
\mathbb{P}(H = 1|I = 0, V = 1) &= 0.01 \\
\mathbb{P}(H = 1|I = 0, V = 0) &= 0.02 \\
\mathbb{P}(H = 1|I = 1, V = 1) &= 0.05 \\
\mathbb{P}(H = 1|I = 1, V = 0) &= 0.10
\end{aligned}$$

Combinando essas probabilidades, temos então que

$$\begin{aligned}
\mathbb{P}(H = 1|V = 1) &= \mathbb{P}(H = 1|V = 1, I = 0)\mathbb{P}(I = 0|V = 1) \\
&\quad + \mathbb{P}(H = 1|V = 1, I = 1)\mathbb{P}(I = 1|V = 1) \\
&\approx 0.01 \times 0.357 + 0.05 \times 0.643 = 0.03572. \qquad (1)
\end{aligned}$$

Da mesma forma, temos que

$$\mathbb{P}(H = 1|V = 0) \approx 0.02 \times 0.833 + 0.1 \times 0.167 = 0.03336. \qquad (2)$$

Concluímos das Equações 1 e 2 que $\mathbb{P}(H = 1|V = 1) > \mathbb{P}(H = 1|V = 0)$. Isso significa que é melhor não se vacinar? Não! Essa desigualdade apenas mostra que, na população estudada, vacinados têm mais chance de serem hospitalizados que não-vacinados. Isto ocorre pois a maior parte dos vacinados são idosos, que têm maior chance de hospitalização que jovens mesmo quando vacinados. Mas a vacina diminui a chance de hospitalização tanto em idosos quanto em jovens. Isso demonstra que avaliar (ou estimar, no caso em que apenas temos uma amostra) $\mathbb{P}(Y = \text{hospitalização}|X = \text{vacinado})$ não responde a pergunta que temos interesse: *se eu aplicasse a vacina a todos os indivíduos da população, qual fração ficaria hospitalizada?*

■

Para escrever formalmente a pergunta que temos interesse, necessitamos introduzir uma linguagem adicional. Utilizaremos para isso o operador *do* (faz). A operação $\text{do}(V = 1)$ é entendida como um *intervenção* em um indivíduo: *"forçamos"* (em um *thought-experiment*) esse indivíduo a tomar a vacina. Assim, no Exemplo 1, desejamos calcular $\mathbb{P}(H = 1|\text{do}(V = 1))$. Em geral, temos interesse em calcular

$$\mathbb{P}(Y \in A|\text{do}(X = x))$$

para diversos eventos A de interesse. Existem duas principais linguagens para inferência causal; cada uma define matematicamente o operador *do* de maneira diferente: a linguagem dos *Potential*

[1] Aqui fizemos uma simplificação para torná-lo mais didático; veja o exemplo original em https://tinyurl.com/simpsoncovid.

Outcomes (possível saídas) (Rubin, 2005) e a linguagem dos Structural Causal Models (modelos causais estruturais) (Pearl, 2009), que faz uso de grafos causais. Ambas as teorias apresentam diferentes perspectivas sobre o modelo problema, e frequentemente podem ser usadas complementarmente para fornecer diferentes *insights*. De fato, elas são matematicamente equivalentes.

Neste texto assumiremos que conhecemos a estrutura causal do problema que estamos resolvendo. Mais especificamente, assumiremos que sabemos (i) quais são as variáveis relevantes para ele (tanto as presentes quanto as ausentes no banco de dados) e (ii) como elas se relacionam causalmente (ou seja, o que causa o que). Essa estrutura é um conjunto de suposições fortes que, na maioria das vezes, não podem ser completamente testadas[2].

Para resolver um problema de inferência causal, após elicitar sua estrutura causal, devemos verificar se o efeito que desejamos estimar é ou não *identificável*. Em palavras, o efeito causal é identificável se ele pode ser estimado a partir dos dados medidos. Muitas vezes isso não é possível; neste caso, dizemos que o problema é *não identificável*. As ferramentas vistas a seguir nos ajudam a verificar se, e como, um efeito causal é identificado. Note que o processo de identificação não depende dos dados, apenas de nossas suposições (embora dados possam auxiliar em alguns aspectos).

2 Potential Outcomes

Por simplicidade, vamos assumir que $X \in \{0, 1\}$ (chamaremos aqui essa variável de *tratamento*). Denotaremos por $Y(1)$ o valor da variável resposta nesse indivíduo caso ele seja submetido a $X = 1$ e $Y(0)$ o valor caso ele seja submetido a $X = 0$. Assim, cada indivíduo tem associado a ele um par $(Y(0), Y(1))$, chamado de *potential outcomes*. Contudo, somente um Y é observado na prática:

$$Y = Y(0)\mathbb{I}(\text{se } X = 0) + Y(1)\mathbb{I}(\text{se } X = 1).$$

A variável não observada é chamada de contrafatual. Com essa nomenclatura, $\mathbb{P}(Y \in A|\text{do}(X = x))$ é definido como $\mathbb{P}(Y(x) \in A)$.

Em geral, **não** é razoável assumir que X é independente do par $(Y(0), Y(1))$. Por exemplo, no Exemplo 1, um indivíduo vacinado $(X = 1)$ tem maior probabilidade de ser idoso, de modo que ele possivelmente tem mais probabilidade de ser hospitalizado (com ou sem vacina). Em outras palavras, esperamos que a distribuição de $(Y(0), Y(1))|X = 1$ seja diferente da distribuição de $(Y(0), Y(1))|X = 0$. Isso implica que, em geral,

$$\begin{aligned}\mathbb{P}(Y \in A|\text{do}(X = 1)) &:= \mathbb{P}(Y(1) \in A) \\ &\neq \mathbb{P}(Y(1) \in A|X = 1) = \mathbb{P}(Y \in A|X = 1),\end{aligned}$$

de modo que as ferramentas usuais de regressão (que visam estimar $\mathbb{P}(Y \in A|X = x)$) não fornecerão uma estimativa do efeito causal. Contudo, é possível estimar $\mathbb{P}(Y \in A|\text{do}(X = x))$ em algumas situações. Veremos aqui duas formas de se fazer isso: (i) ajustar por variáveis de confusão e (ii) desenhar um experimento aleatorizado. Existem, contudo, outras maneiras de estimar o efeitos causais em alguns casos como, por exemplo, quando variáveis instrumentais estão disponíveis (Alves, 2021).

2.1 Ajuste por variáveis de confusão

Uma forma de tornar o efeito causal estimável é medir *variáveis de confusão*, que são variáveis que afetam tanto X quanto Y. Seja $\mathbf{Z}$ um vetor de variáveis de confusão. Se assumirmos que todos os confundidores foram medidos (uma hipóteses frequentemente chamada de *ignorabilidade*), conseguimos medir o efeito causal a partir de apenas dados observacionais (isto é, não é necessário desenhar um experimento aleatorizado, como descrito na Seção 2.2).

[2]Existe, contudo, uma extensa literatura sobre como descobrir estruturas causais (Glymour et al., 2019).

Suposição 1 (Todos confundidores medidos)**.**

$$X \text{ é independente de } (Y(0), Y(1)) \text{ condicionalmente em } \mathbf{Z}.$$

Para interpretar a Suposição 1, vamos assumir, por simplicidade, que Y indica que um indivíduo se recuperou ($Y = 1$) ou não ($Y = 0$) de uma doença. Segundo essa suposição, $\mathbb{P}(Y(0) = 1|\mathbf{Z} = \mathbf{z}, X = 0) = \mathbb{P}(Y(0) = 1|\mathbf{Z} = \mathbf{z}, X = 1)$. Assim, ela implica que a probabilidade de um indivíduo com covariáveis $\mathbf{Z} = \mathbf{z}$ que não foi submetido ao tratamento se recuperar sem ele é a mesma caso ele tivesse sido alocado para receber o tratamento. Analogamente, essa suposição também equivale a $\mathbb{P}(X = 1|\mathbf{Z} = \mathbf{z}, (Y(0), Y(1))) = \mathbb{P}(X = 1|\mathbf{Z} = \mathbf{z})$, isto é, dado $\mathbf{z}$, a probabilidade de um indivíduo ser tratado não depende das saídas $(Y(0), Y(1))$. Assim, no Exemplo 1, assumir $V \perp\!\!\!\perp (H(0), H(1))|I$ implica que, dada a idade de um indivíduo, a probabilidade dele tomar a vacina não depende dos seus potenciais status de hospitalização $(H(0), H(1))$.

O seguinte teorema mostra como o efeito causal se relaciona com a distribuição geradora dos dados e, assim, indica como ele pode ser estimado a partir de dados observacionais.

Teorema 1. *Sob a Suposição 1,*

$$\mathbb{P}(Y \in A|do(X = x)) = \int \mathbb{P}(Y \in A|X = x, \mathbf{Z} = \mathbf{z}) d\mathbb{P}(\mathbf{z}).$$

Demonstração.

$$\begin{aligned}
\mathbb{P}(Y \in A|\text{do}(X = x)) &= \mathbb{P}(Y(x) \in A) && \text{Definição} \\
&= \int \mathbb{P}(Y(x) \in A|\mathbf{Z} = \mathbf{z}) d\mathbb{P}(\mathbf{z}) && \text{Lei da probabilidade total} \\
&= \int \mathbb{P}(Y(x) \in A|X = x, \mathbf{Z} = \mathbf{z}) d\mathbb{P}(\mathbf{z}) && \text{Suposição 1} \\
&= \int \mathbb{P}(Y \in A|X = x, \mathbf{Z} = \mathbf{z}) d\mathbb{P}(\mathbf{z}). && \text{Definição}
\end{aligned}$$

■

O termo $\int \mathbb{P}(Y \in A|X = x, \mathbf{Z} = \mathbf{z}) d\mathbb{P}(\mathbf{z})$ pode ser estimado com base em um conjunto de dados, pois é uma função da distribuição conjunta dos dados. Exploraremos isso nos exemplos e exercícios.

Exemplo 2. **[Continuação do Exemplo 1]** Vamos assumir que a única variável de confusão é a idade (Suposição 1). Segue do Teorema 1 que

$$\mathbb{P}(H = 1|\text{do}(V = 1)) = \sum_{i=0}^{1} \mathbb{P}(H = 1|V = 1, I = i)\mathbb{P}(I = i) = 0.01 \times 0.5 + 0.05 \times 0.5 = 0.03.$$

Analogamente,

$$\mathbb{P}(H = 1|\text{do}(V = 0)) = \sum_{i=0}^{1} \mathbb{P}(H = 1|V = 0, I = i)\mathbb{P}(I = i) = 0.02 \times 0.5 + 0.1 \times 0.5 = 0.06,$$

de modo que $\mathbb{P}(H = 1|\text{do}(V = 0)) > \mathbb{P}(H = 1|\text{do}(V = 1))$ (é melhor se vacinar!).

■

É frequente termos interesse em funções de $\mathbb{P}(Y \in A|\text{do}(X = x))$. Por exemplo, em um contexto de regressão (isto é, $Y \in \mathbb{R}$), o *efeito causal médio* (ATE; *average treatment effect*), é definido como

$$\mathbb{E}[Y|\text{do}(X = 1)] - \mathbb{E}[Y|\text{do}(X = 0)].$$

Essa quantidade pode ser estimada fazendo a diferença entre duas funções de regressão.

Exemplo 3 (Estimação do ATE)**.** Seja $\mu(x, \mathbf{z}) := \mathbb{E}[Y|x, \mathbf{z}]$ a função de regressão de Y em $(x, \mathbf{z})$. Segue do Teorema 1 que

$$\begin{aligned}\mathbb{E}[Y|\text{do}(X=1)] - \mathbb{E}[Y|\text{do}(X=0)] &= \int \mathbb{E}(Y|X=1, \mathbf{Z}=\mathbf{z}) d\mathbb{P}(\mathbf{z}) - \int \mathbb{E}(Y|X=0, \mathbf{Z}=\mathbf{z}) d\mathbb{P}(\mathbf{z}) \\ &= \int \mu(1, \mathbf{z}) d\mathbb{P}(\mathbf{z}) - \int \mu(0, \mathbf{z}) d\mathbb{P}(\mathbf{z}).\end{aligned}$$

Assim, o ATE pode ser estimado via

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}(1, \mathbf{z}_i) - \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}(0, \mathbf{z}_i), \tag{3}$$

em que $\widehat{\mu}(x, \mathbf{z})$ é uma estimativa da regressão $\mu(x, \mathbf{z})$ obtida usando uma amostra aleatória $(X_1, \mathbf{Z}_1, Y_1), \ldots, (X_n, \mathbf{Z}_n, Y_n)$.

■

O exemplo a seguir desenvolve esse estimador do ATE para o caso de um modelo linear.

Exemplo 4 (Modelo Linear)**.** Suponha que a função de regressão seja linear, isto é, $\mu(x, \mathbf{z}) = \beta_0 + \beta_x x + \beta_{\mathbf{z}}^t \mathbf{z}$. Neste caso, segue do Exemplo 3 que

$$\begin{aligned}\mathbb{E}[Y|\text{do}(X=1)] - \mathbb{E}[Y|\text{do}(X=0)] &= \int \mu(1, \mathbf{z}) d\mathbb{P}(\mathbf{z}) - \int \mu(0, \mathbf{z}) d\mathbb{P}(\mathbf{z}) \\ &= \int (\beta_0 + \beta_x \times 1 + \beta_{\mathbf{z}}^t \mathbf{z}) d\mathbb{P}(\mathbf{z}) - \int (\beta_0 + \beta_x \times 0 + \beta_{\mathbf{z}}^t \mathbf{z}) d\mathbb{P}(\mathbf{z}) \\ &= \beta_x.\end{aligned}$$

Em palavras, o coeficiente da regressão de Y em $(X, \mathbf{Z})$ associado a X é uma estimativa do efeito causal de X em Y desde que todas as variáveis confundidoras foram de fato medidas. Assim, podemos fazer um ajuste via mínimos quadrados para estimar $\mathbb{E}[Y|x, \mathbf{z}]$ e utilizar $\widehat{\beta}_x$ como estimativa do ATE de X em Y.

■

Exercício 1 (Regressão linear com interações)**.** Faça um desenvolvimento similar ao do Exemplo 4, mas agora assumindo que o modelo linear inclui as interações entre x e cada componente de $\mathbf{z}$.

■

Exercício 2 (Regressão Logística)**.** Considere que $Y \in \{0, 1\}$ e que $\mathbb{P}(Y = 1|x, \mathbf{z})$ é modelada por uma regressão logística. Derive estimadores para as seguintes quantidades:

- Causal difference risk: $\mathbb{P}(Y = 1|\text{do}(X = 1)) - \mathbb{P}(Y = 1|\text{do}(X = 0)$
- Causal risk ratio: $\frac{\mathbb{P}(Y=1|\text{do}(X=1))}{\mathbb{P}(Y=1|\text{do}(X=0))}$
- Causal odds ratio: $\frac{\frac{\mathbb{P}(Y=1|\text{do}(X=1))}{\mathbb{P}(Y=0|\text{do}(X=1))}}{\frac{\mathbb{P}(Y=1|\text{do}(X=0))}{\mathbb{P}(Y=0|\text{do}(X=0))}}$

■

Exercício 3 (Regressão Linear - X Quantitativo)**.** Mostre como a Suposição 1 deve ser adaptada para o caso em que $X \in \mathbb{R}$ e mostre como definir o ATE neste caso. Mostre também quanto ele vale no caso de um modelo linear.

■

2.2 Aleatorização

Embora a Seção 2.1 mostre uma forma de estimar o efeito causal com base em estudos observacionais, esta abordagem possui várias limitações na prática:

- Nem sempre sabemos quais são os confundidores
- Nem sempre podemos medir todos os confundidores
- É fundamental que o modelo usado para estimar a função de regressão seja corretamente especificado

Uma forma alternativa de tornar o efeito causal estimável com menos suposições é aleatorizar X. Isto é, definiremos $X \sim \text{Ber}(p)$ para algum $0 < p < 1$. Com isso, por construção teremos que X é independente de $(Y(0), Y(1))$, de modo que

$$\mathbb{P}(Y \in A|\text{do}(X = x)) = \mathbb{P}(Y(x) \in A) = \mathbb{P}(Y(x) \in A|X = x) = \mathbb{P}(Y \in A|X = x),$$

isto é, o efeito causal é exatamente $\mathbb{P}(Y \in A|X = x)$, que podemos estimar com métodos usuais.

Ao contrário do que métodos baseados em controle por confundidores, aleatorizar requer menos suposições para derivar efeitos causais. Contudo, nem sempre é possível aleatorizar (tanto por restrições físicas quanto por restrições éticas).

3 Grafos Causais

Neste abordagem, representamos nosso conhecimento causal sobre as variáveis envolvidas em um problema através de um grafo[3]. A partir desse grafo, podemos derivar quais variáveis devem ser incluídas no fórmula do ajuste (que será novamente derivada aqui utilizando essa nova nomenclatura).

3.1 Grafos Direcionados Acíclicos (DAGs)

Um grafos direcionado acíclico (DAG) é uma representação gráfica da estrutura de dependência de um vetor aleatório $\mathbf{W} := (W_1, \ldots, W_d)$ em que cada vértice representa uma variável. Veja um exemplo na Figura 2. Em um DAG, se há uma seta de W_i até W_j, dizemos que W_i é pai de W_j. Denotamos por $\pi(W_i)$ o conjunto de todos os pais do vértice W_i. Para que um DAG represente adequadamente a estrutura de dependência de $\mathbf{W}$, a seguinte decomposição deve valer:

$$p(\mathbf{w}) = \prod_{i=1}^{d} p(w_i|\pi(w_i)).$$

Um DAG permite que diversas independências condicionais sejam derivadas facilmente. Por exemplo, em um DAG, temos que

$$W_i \text{ é independente de } \mathbf{W}' \text{ condicionalmente em } \pi(W_i), \tag{4}$$

em que $\mathbf{W}'$ representa todas as demais variáveis além de $W_i, \pi(W_i)$ e os descendentes de W_i. Por exemplo, o DAG da Figura 2 implica que $W_4 \perp\!\!\!\perp W_1|W_3$ e $W_2 \perp\!\!\!\perp W_3|W_1$ (entre outros).

[3] Grafos causais são um componente de uma abordagem chamada Structural Causal Models; veja (Pearl, 2009).

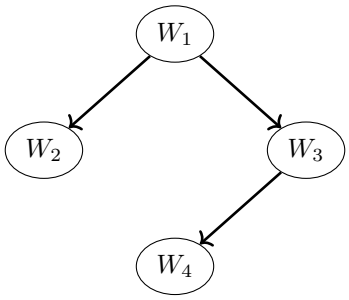


Figura 2: Exemplo de DAG.

Exercício 4. Derive três independências condicionais implicadas pelo DAG da Figura 3.

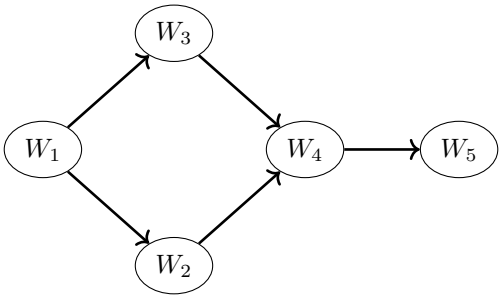


Figura 3: DAG do Exercício 4.

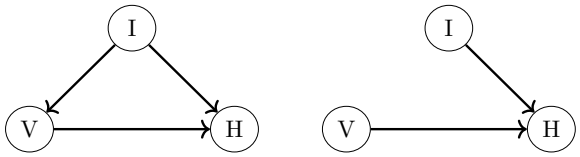
■

Grafos têm muitas aplicações em estatística não relacionadas com causalidade. Para que um DAG também seja útil em inferência causal, devemos atribuir a ele uma interpretação causal. Chamaremos um DAG de um *DAG causal* se ele codificar corretamente efeitos causais, no sentido que se há uma flecha de W_i a W_j, então W_i é uma causa direta de W_j.[4] A Figura 4 (esquerda) ilustra um DAG causal que pode ser associado ao Exemplo 1. Ao explicitar um grafo causal, assumimos que todas as variáveis relevantes ao problema estão inclusas nele. Por exemplo, se também achamos que a variável sexo é tal que Vacina ← Sexo → Hospitalização, então ela deve ser incluída no modelo para que as conclusões causais sejam corretas.

Dado um DAG causal, o efeito de uma intervenção pode ser representado através da remoção de algumas flechas. Mais especificamente, representamos o efeito da intervenção $\text{do}(X = x)$ removendo todas as flechas que chegam em X. A Figura 4 (direita) ilustra o efeito de uma intervenção em vacina no Exemplo 1. Assim, a intervenção induz uma nova medida de probabilidade para as variáveis do grafo, que denotaremos por $\mathbb{P}_{\text{do}(X=x)}$, que é usada para definir o efeito causal. Nesta abordagem, o efeito causal de X em Y é definido através de

$$\mathbb{P}(Y \in A|\text{do}(X = x)) := \mathbb{P}_{\text{do}(X=x)}(Y \in A|X = x).$$

[4]Dizemos que W_i é uma causa de W_j se W_i é uma causa direta de W_j, ou de alguma outra causa de W_j.



(a) DAG original

(b) DAG após a intervenção em vacina

Figura 4: Esquerda: grafo original para o Exemplo 1. Direita: grafo após a intervenção na variável "Vacina".

No contexto de DAGs, um vetor $\mathbf{Z}$ é um vetor de confusão possui a estrutura do DAG da Figura 5. Neste caso, o efeito causal de X em Y pode ser facilmente calculado utilizando a correção dada pelo seguinte teorema.

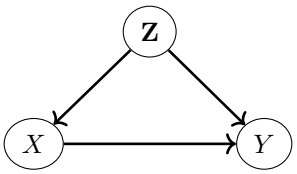


Figura 5: Neste DAG, $\mathbf{Z}$ confunde o efeito de X em Y.

Teorema 2. *Sob o DAG da Figura 5,*

$$\mathbb{P}(Y \in A|do(X = x)) = \int \mathbb{P}(Y \in A|X = x, \mathbf{Z} = \mathbf{z})d\mathbb{P}(\mathbf{z}).$$

Demonstração. Primeiramente, note que $p(x, y, \mathbf{z}) = p(\mathbf{z})p(x|\mathbf{z})p(y|x, \mathbf{z})$ e $p_{\text{do}(X=x_0)}(x, y, \mathbf{z}) = p(\mathbf{z})\mathbb{I}(x = x_0)p(y|x, \mathbf{z})$. Assim,

$$\begin{aligned}
\mathbb{P}(Y \in A|\text{do}(X = x)) &= \mathbb{P}_{\text{do}(X=x)}(Y \in A|X = x) && \text{Definição} \\
&= \int \mathbb{P}_{\text{do}(X=x)}(Y \in A|X = x, \mathbf{Z} = \mathbf{z})d\mathbb{P}_{\text{do}(X=x)}(\mathbf{z}|X = x) && \text{Lei da probabilidade total} \\
&= \int \mathbb{P}_{\text{do}(X=x)}(Y \in A|X = x, \mathbf{Z} = \mathbf{z})d\mathbb{P}_{\text{do}(X=x)}(\mathbf{z}). && \text{DAG da Fig. 5 após intervenção} \\
&= \int \mathbb{P}(Y \in A|X = x, \mathbf{Z} = \mathbf{z})d\mathbb{P}(\mathbf{z}). && \text{Invariância dessas } \mathbb{P}\text{'s nos DAGs}
\end{aligned}$$

■

Assim, a formula de ajuste dado pela teoria de DAGs (Teorema 2) é a mesma que aquela apresentada no contexto de *potential outcomes* (Teorema 1).

Exercício 5. Argumente, na linguagem de DAGs, porque aleatorizar X faz com que seja possível estimar o efeito causal de X em Y diretamente.

■

3.1.1 Quais variáveis controlar?

A derivação que fizemos da fórmula do ajuste (Teorema 2) assume que a relação de $\mathbf{Z}$ com X e Y é dada pelo DAG da Figura 5. Contudo, nem toda variável é uma variável de confusão.

Mostraremos, agora, que incluir uma variável que não é de confusão na fórmula do ajuste (Teorema 2) pode ser extremamente perigoso, no sentido que fórmula do ajuste não necessariamente retorna $\mathbb{P}(Y|\text{do}(X = x))$ se incluímos essas variáveis nela. Ilustraremos esse fenômeno aqui com um tipo de variável que é particularmente danosa caso seja incluída na fórmula do ajuste: um *collider*. $\mathbf{Z}$ é um collider (colisor) se ele possui a estrutura dada pelo DAG da Figura 6. Informalmente, $\mathbf{Z}$ é um collider se ele é consequência tanto de X quanto de Y. O exemplo a seguir mostra que não devemos condicionar em colliders ao se estimar o efeito de X em Y.

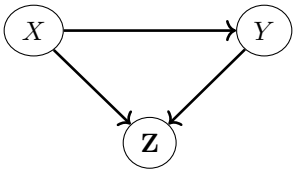


Figura 6: Neste DAG, $\mathbf{Z}$ é um collider.

Exemplo 5. [Collider no exemplo da vacina.] Consideremos agora uma modificação do Exemplo 1. Nesta versão, aplicamos uma vacina que **não** foi desenhada para COVID, mas sim para combater uma outra doença pega frequentemente em ambientes hospitalares (mas não apenas neles). Seja $D = \mathbb{I}(\text{doença})$ a variável que indica se um indivíduo contraiu essa doença no período do estudo. Assim, as suposições causais feitas aqui são dadas pelo DAG da Figura 7.

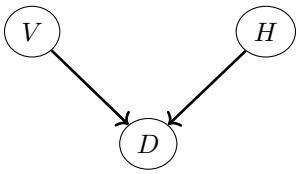


Figura 7: DAG do Exemplo 5. D representa uma doença que é evitada pela vacina, mas que pode ser causada por hospitalizações por outras doenças.

Segundo este DAG, $\mathbb{P}(H|\text{do}(V = 1)) = \mathbb{P}(H)$, de modo que de fato ele implica que a vacina não é eficaz contra hospitalizações por COVID. Contudo, se usamos a fórmula da correção (Teorema 2) como se D fosse uma variável de confusão, poderíamos checar à conclusão que a vacina é eficaz. De fato, se $\mathbb{P}(H = 1) = 1/2$, $\mathbb{P}(V = 1) = 0.1$, $\mathbb{P}(D = 1|H = 0, V = 0) = 0.1$, $\mathbb{P}(D = 1|H = 1, V = 0) = 0.5$, $\mathbb{P}(D = 1|H = 0, V = 1) = 0$ e $\mathbb{P}(D = 1|H = 1, V = 1) = 0.1$, temos que a formula da correção que usa D como variável de confusão indica que

$$\mathbb{P}(H = 1|D = 1, V = 1)\mathbb{P}(D = 1) + \mathbb{P}(H = 1|D = 0, V = 1)\mathbb{P}(D = 0) \approx 0.62.$$

Se fosse apropriado usar essa fórmula para calcular $\mathbb{P}(H|\text{do}(V = 1))$, concluiríamos então que $\mathbb{P}(H|\text{do}(V = 1)) > \mathbb{P}(H)$ (ou seja, que a vacina previne hospitalizações por COVID), o que é uma contradição com o DAG.

■

Exercício 6. [Mediação.] Neste exercício, veremos outro tipo de variável que não deve ser usada na fórmula de ajuste, um *mediador*. Para isso, voltemos ao exemplo da vacina. Desta vez, vamos considerar além de V (vacinação) e H (hospitalização), a variável M (a variável que indica se um indivíduo usa máscaras). Vamos assumir o DAG da Figura 8. Em palavras, vamos considerar que um individuo ser vacinado por fazer com que ele deixe de usar máscaras (por exemplo), e que máscaras também mudem a chance de hospitalização. Formalmente, M *media* o efeito de V

em H. Argumente que, nesse caso, a fórmula do ajuste com M não retorna, necessariamente, $\mathbb{P}(H \in A|\text{do}(V = v))$. Qual passo da demonstração deste teorema falha?

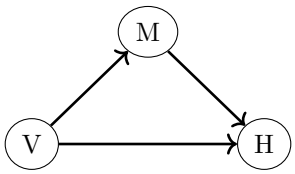


Figura 8: DAG do Exemplo 6.

■

Há casos em que condicionar em algumas variáveis não é necessário, mas também não invalida a análise (no sentido de que se elas forem adicionadas à fórmula do ajuste, essa ainda retornará $\mathbb{P}(Y \in A|\text{do}(X = x))$), como mostra o seguinte exercício.

Exercício 7. [Controle por outras variáveis explicativas.] Dê um exemplo de problema para o qual o DAG da Figura 9 seja compatível. Mostre que, sob esse DAG, utilizar a fórmula do ajuste controlando por $(\mathbf{Z}, W)$ retorna $\mathbb{P}(Y \in A|\text{do}(X = x))$. Mostre o mesmo controlando apenas por $\mathbf{Z}$. Conclua que ambas as escolhas são válidas. (Controlar por $(\mathbf{Z}, W)$, contudo, por levar a uma diminuição da variância da estimativa do efeito causal; Cinelli et al. 2021.)

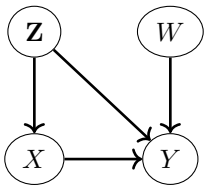


Figura 9: DAG do Exercício 7.

■

Veja Cinelli et al. (2021) para diversos outros exemplos em DAGs mais complexos que os estudados aqui. Veja também Textor et al. (2016)[5] para uma ferramenta simples que automaticamente calcula quais variáveis devem ser incluídas na fórmula do ajuste para que seja possível estimar um efeito causal em DAGs mais complexos. A matemática por trás dessas derivações é chamada de *do-calculus* e pode ser encontrada em Pearl (2009).

4 Inverse Probability Weighting e Propensity Scores

A fórmula do ajuste (Teoremas 1 e 2) nos fornece uma maneira de estimar o efeito causal de X em Y caso seja possível medir todos os confundidores $\mathbf{Z}$. Em outras palavras, ela mostra como o efeito causal se relaciona matematicamente à distribuição do dados observáveis. Essencialmente, ele mostra que basta adicionar os confundidores $\mathbf{Z}$ como covariáveis na regressão de Y em X para se obter uma estimativa do efeito causal. Contudo, para que esse efeito seja bem estimado, é necessário que o modelo de regressão seja corretamente especificado. Algumas maneiras alternativas para estimar efeitos causais na presença de confundidores podem ser criadas a partir do escore de propensão (*propensity score*), que estudamos nessa seção.

[5] http://www.dagitty.net/dags.html#

Por simplicidade, vamos assumir que o tratamento $X \in \{0,1\}$. O escore de propensão é a função

$$e(\mathbf{z}) := \mathbb{P}(X = 1|\mathbf{z}).$$

Em palavras, essa função mede a probabilidade de um individuo ser tratado dadas as suas variáveis de confusão $\mathbf{z}$. Em estudos observacionais, $e(\mathbf{z})$ não pode ser calculado, mas, caso se as variáveis de confusão sejam medidas, pode-se estimá-lo (por exemplo, com uma regressão logística de X em $\mathbf{Z}$). A seguir veremos duas maneiras de utilizar esse escore.

4.1 Inverse Probability Weighting (IPW)

Uma forma de se usar escores de propensão é dada pelo seguinte teorema.

Teorema 3. *Sob a Suposição 1,*

$$\mathbb{E}[Y|do(X = 1)] = \mathbb{E}\left[\frac{Y\mathbb{I}(X = 1)}{e(\mathbf{Z})}\right].$$

Demonstração.

$$\begin{aligned}
\mathbb{E}[Y|\text{do}(X = 1)] &= \int \mathbb{E}(Y|X = 1, \mathbf{Z} = \mathbf{z})dP(\mathbf{z}) && \text{Teorema 1}\\
&= \int\int yp(y|\mathbf{z}, X = 1)dydP(\mathbf{z})\\
&= \int\int y\frac{p(y, \mathbf{z}, X = 1)}{e(\mathbf{z})}dyd\mathbf{z}\\
&= \int\int \sum_{x=0}^{1} y\mathbb{I}(x = 1)\frac{p(y, \mathbf{z}, x)}{e(\mathbf{z})}dyd\mathbf{z}\\
&= \mathbb{E}\left[\frac{Y\mathbb{I}(X = 1)}{e(\mathbf{Z})}\right] && \text{Lei do estatístico inconsciente}
\end{aligned}$$

■

O Teorema 3 indica que se $\widehat{e}(\mathbf{z})$ é uma estimativa do propensity score, então pode-se estimar o ATE via

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i\mathbb{I}(X_i = 1)}{\widehat{e}(\mathbf{Z}_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i\mathbb{I}(X_i = 0)}{1 - \widehat{e}(\mathbf{Z}_i)}, \tag{5}$$

em que $(X_1, \mathbf{Z}_1, Y_1), \ldots, (X_n, \mathbf{Z}_n, Y_n)$ é uma amostra aleatória. Idealmente, a amostra usada para calcular essa estimativa deve ser diferente da amostra usada para estimar $e(\mathbf{z})$. Para que a Equação 5 seja uma boa estimativa, o propensity score deve estar bem estimado. Além disso, $\widehat{e}(\mathbf{Z}_i)$ não pode ficar próximo de zero nem um, caso contrário a estimativa do ATE terá uma variabilidade muito grande.

4.1.1 Estimador duplamente robusto

A ideia do estimador duplamente robusto é combinar os estimadores das Equações 3 e 5 de modo a se obter um estimador mais robusto. Mais precisamente, ele é desenhado de forma que tenha bom desempenho se o modelo para $\mu(x, \mathbf{z})$ ou $e(\mathbf{z})$ for razoável (ou seja, ele é robusto ao caso de um dos modelos ser ruim).

O estimador duplamente robusto do ATE é definido por

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathbb{I}(X_i = 1)\,(Y_i - \widehat{\mu}(1, \mathbf{Z}_i))}{\widehat{e}(\mathbf{Z}_i)} + \widehat{\mu}(1, \mathbf{Z}_i)\right] - \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathbb{I}(X_i = 0)\,(Y_i - \widehat{\mu}(0, \mathbf{Z}_i))}{1 - \widehat{e}(\mathbf{Z}_i)} + \widehat{\mu}(0, \mathbf{Z}_i)\right]. \tag{6}$$

Para notar que esse estimador de fato é robusto, observe que:

- Se a estimativa de $\mu(x, \mathbf{z})$ for boa, então

$$\mathbb{E}\left[\frac{\mathbb{I}(X_i = 1)\,(Y_i - \widehat{\mu}(1, \mathbf{Z}_i))}{\widehat{e}(\mathbf{Z}_i)}\right] \approx 0 \text{ e } \mathbb{E}\left[\frac{\mathbb{I}(X_i = 0)\,(Y_i - \widehat{\mu}(0, \mathbf{Z}_i))}{1 - \widehat{e}(\mathbf{Z}_i)}\right] \approx 0.$$

  Assim, pela lei dos grandes números, segue que

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{I}(X_i = 1)\,(Y_i - \widehat{\mu}(1, \mathbf{Z}_i))}{\widehat{e}(\mathbf{Z}_i)} \approx 0 \text{ e } \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{I}(X_i = 0)\,(Y_i - \widehat{\mu}(0, \mathbf{Z}_i))}{1 - \widehat{e}(\mathbf{Z}_i)} \approx 0$$

  e, assim, o estimador duplamente robusto é aproximadamente o estimador da Equação 3. Ou seja, caso μ seja bem estimado, o estimador não usará os propensity scores.

- Se a estimativa de $e(\mathbf{z})$ for boa, então

$$\mathbb{E}\left[\frac{\mathbb{I}(X_i = 1) - \widehat{e}(\mathbf{Z}_i)}{\widehat{e}(\mathbf{Z}_i)}\mu(1, \mathbf{Z}_i)\right] \approx 0 \text{ e } \mathbb{E}\left[\frac{\widehat{e}(\mathbf{Z}_i) - \mathbb{I}(X_i = 1))}{1 - \widehat{e}(\mathbf{Z}_i)}\mu(0, \mathbf{Z}_i)\right] \approx 0.$$

  Como a Equação 6 pode ser reescrita como

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathbb{I}(X_i = 1)Y_i}{\widehat{e}(\mathbf{Z}_i)} - \frac{\mathbb{I}(X_i = 1) - \widehat{e}(\mathbf{Z}_i)}{\widehat{e}(\mathbf{Z}_i)}\widehat{\mu}(1, \mathbf{Z}_i)\right] - \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathbb{I}(X_i = 0)Y_i}{1 - \widehat{e}(\mathbf{Z}_i)} - \frac{\widehat{e}(\mathbf{Z}_i) - \mathbb{I}(X_i = 1)}{1 - \widehat{e}(\mathbf{Z}_i)}\widehat{\mu}(0, \mathbf{Z}_i)\right],$$

  concluímos que, nesse caso, o estimador duplamente robusto é aproximadamente da Equação 5. Ou seja, caso e seja bem estimado, o estimador não usará $\widehat{\mu}$.

4.2 Propensity score como redução de dimensão

O propensity score pode ser visto como uma transformação das variáveis de confusão $\mathbf{Z}$ para $[0, 1]$. O seguinte teorema mostra que essa transformação ainda permite que o efeito causal de X em Y seja estimado.

Teorema 4. *Sob a Suposição 1,*

$$X \text{ é independente de } (Y(0), Y(1)) \text{ condicionalmente em } e(\mathbf{Z}).$$

Segue dos Teoremas 1 e 4 que

$$\mathbb{P}(Y \in A | \text{do}(X = x)) = \int \mathbb{P}(Y \in A | X = x, e(\mathbf{Z}) = e) d\mathbb{P}(e).$$

Em particular, esse resultado implica que podemos estimar o ATE fazendo uma regressão de Y no par $(x, e(\mathbf{z}))$. Isto é, não é necessário adicionar $\mathbf{z}$ como um todo. Claro, na prática temos primeiro que estimar $e(\mathbf{z})$. Esse resultado também sugere outras formas de usar o propensity score, como *matching* (Rosenbaum and Rubin, 1983), que não exploramos aqui.

4.3 Propensity scores em estudos aleatorizados

Muitas vezes, podemos fazer um estudo aleatorizado em que cada indivíduo não tem a mesma probabilidade de pertencer ao grupo que recebe o tratamento. Essas probabilidades são pré-determinadas a partir de covariáveis $\mathbf{z}$. Neste contexto, o propensity score $e(\mathbf{z})$ é conhecido, e podemos usar as ferramentas descritas acima para estimar o efeito causal sem precisar estimar $e(\mathbf{z})$.

Exercício 8. Considere o banco de dados `synthetic_data.csv`, disponível em https://github.com/grf-labs/grf/tree/master/experiments/acic18. Esses dados são inspirados no The National Study of Learning Mindsets. Neste estudo, alguns estudantes de escolas públicas nos Estados Unidos foram convidados a assistir uma palestra. O objetivo do estudo é avaliar se essa palestra foi capaz de alterar a mentalidade dessas pessoas de modo a melhorar seu desempenho acadêmico. Responda às seguintes perguntas:

- Quais das variáveis disponíveis você acha que são variáveis de confusão?
- Estime o ATE usando um modelo de regressão linear simples (sem confundidoras).
- Estime o ATE usando um modelo de regressão linear simples, adicionando as variáveis que você acredita ser de confusão.
- Estime o propensity score usando uma regressão logística. Mostre a distribuição desse score nos dois tratamentos. Interprete os resultados.
- Estime o ATE usando o Teorema 4.
- Estime o ATE usando uma regressão linear com o propensity score como covariável.
- Estime o ATE usando o estimador duplamente robusto (Equação 6).
- Repita os itens que dependem do propensity score usando uma estimativa por floresta aleatória.
- Como se comparam as estimativas do ATE que você obteve? Qual você utilizaria no problema real? Justifique.

■

5 Exemplo: estimando contrafatuais

Por vezes, temos interesse em estimar contrafatuais. Ou seja, desejamos estimar o que teria ocorrido se uma decisão diferente tivesse sido tomada. Aqui descrevemos uma forma como isso pode ser feita em um contexto bem específico. Assumiremos que temos T séries temporais e que uma intervenção é feita em um dado momento apenas na primeira dessas séries. Desejamos então estimar o que teria acontecido com essa séries *se a intervenção não houvesse sido feita nela*. Para isso, utilizarmos a ideia de controles sintéticos (Abadie et al., 2010), que consiste em utilizar as séries que não sofreram intervenção para modelar o contrafatual da primeira série.

Para tornar o exemplo mais concreto, utilizaremos o modelo usado por Izbicki et al. (2021) para medir quantas hospitalizações foram evitadas até Maio/2021 em em indivíduos maiores de 65 anos por conta da vacina para COVID. Para isso, precisamos estimar quantas hospitalizações teriamos nesse grupo caso ele não tivesse sido vacinado.

Denotaremos por $Y_{t,f}$ o número de internados com COVID na faixa etária f ($f = 1, \dots, F$) no dia t ($t = 1, \dots, T$). Os potential outcomes serão denotados por $Y_{t,f}(v)$, em que $v = 1$ representa esse número caso esse grupo já tenha sido submetido à vacina no instante t, e $v = 0$ indica o mesmo número caso esse grupo não tenha sido submetido à vacina. Vamos assumir que a intervenção (a vacina) no grupo de interesse (sem perda de generalidade, $f = 1$) ocorreu no instante t_0 e que, até o final do estudo, os outros grupos não sofreram efeito dessa intervenção. Seja $\mathbf{Y}_t^{-1}(v) = (Y_{t,2}(v), \dots, Y_{t,F}(v))$. Assumiremos que

$$Y_{t,f}(0) = \beta_0 + \boldsymbol{\beta}^t \mathbf{Y}_t^{-1}(0) + \epsilon_{t,f}, \ t = 1, \dots, T \tag{7}$$

com $(\epsilon_{t,f})_{t,f} \overset{i.i.d.}{\sim} N(0, \sigma^2)$, embora a abordagem de controles sintéticos seja muito mais geral. Como $Y_{t,f} = Y_{t,f}(0)$ para todo f e $t < t_0$, a Equação 7 implica que

$$Y_{t,1} = \beta_0 + \boldsymbol{\beta}^t \mathbf{Y}_t^{-1} + \epsilon_{t,f} \quad t = 1, \dots, T_0 - 1$$

o que sugere que podemos estimar os parâmetros desse modelo ($\beta_0, \boldsymbol{\beta}$ e σ^2) via uma regressão linear por mínimos quadrados com os dados até T_0. Além disso, como $Y_{t,1} = Y_{t,f}(0)$ para todo $t \geq T_0$, *observamos* $y_{t,f}(0)$ para todo $t \geq T_0$, de modo que podemos estimar o contrafatual no grupo 1 via

$$\widehat{Y}_{t,1}(0) := \widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}^t \mathbf{Y}_t^{-1}(0), \ t = T_0, \dots, T.$$

Como medimos $Y_{t,1}(1)$ para $t \geq T_0$, podemos então avaliar $Y_{t,1}(1) - \widehat{Y}_{t,1}(0)$ para ter uma estimativa do efeito que a intervenção teve em cada instante.

Em Izbicki et al. (2021), utilizamos como único grupo sintético a população entre 55 e 62 anos, pois esse grupo ainda não havia recebido a vacina no período analisado e era mais próximo (em termos de comportamento) do grupo idoso, de modo que ele é mais informativo para modelar o comportamento deste. Os resultados deste modelo para o estado de São Paulo encontram-se na Figura 10. Note que o modelo realmente se ajusta bem antes da intervenção.

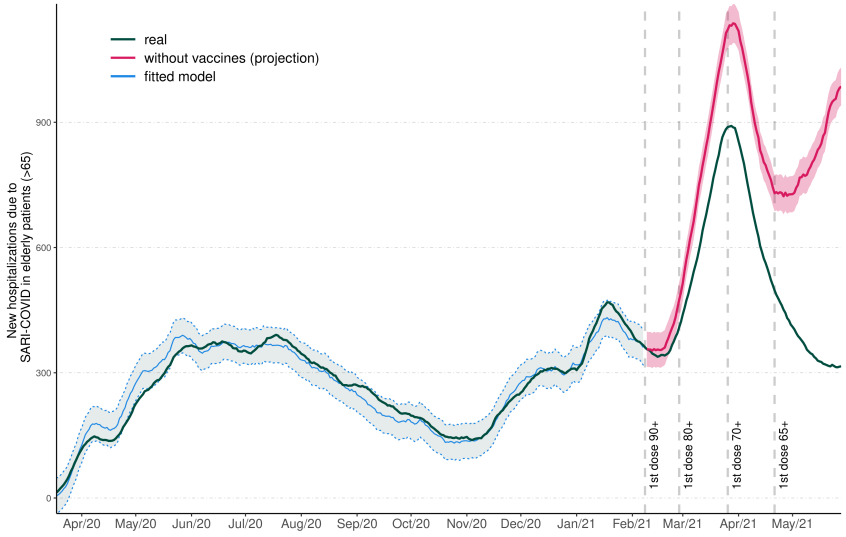


Figura 10: contrafatual estimado para o número de internações evitadas em idosos devido à vacina. Extraído de Izbicki et al. (2021).

Existem muitos outros atalhos para derivar contrafatuais em situações específicas, como DID e Discontinuidade (Alves, 2021).

Referências

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505, 2010.

Matheus Facure Alves. *Causal Inference for The Brave and True*. 2021. URL `matheusfacure.github.io/python-causality-handbook`.

Carlos Cinelli, Andrew Forney, and Judea Pearl. A crash course in good and bad controls. 2021.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Rafael Izbicki and Tiago Mendonça dos Santos. *Aprendizado de máquina: uma abordagem estatística*. 2020.

Rafael Izbicki, Leonardo S Bastos, Meyer Izbicki, Hedibert F Lopes, and Tiago Mendonça dos Santos. How many hospitalizations has the covid-19 vaccination already prevented in são paulo? *Clinics*, 76, 2021.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Johannes Textor, Benito van der Zander, Mark S Gilthorpe, Maciej Liśkiewicz, and George TH Ellison. Robust causal inference using directed acyclic graphs: the r package 'dagitty'. *International journal of epidemiology*, 45(6):1887–1894, 2016.