Inferência Causal

Notas de Aula

Rafael Bassi Stern

Última revisão: 13 de março de 2023

Por favor, enviem comentários, typos e erros para rbstern@gmail.com

Agradecimentos:

"Teaching is giving opportunities to students to discover things by themselves."

George Pólya

Conteúdo

1. Por que estudar Inferência Causal?

Você já deve ter ouvido diversas vezes que **correlação não implica causalidade**. Contudo, o que é causalidade e como ela pode ser usada para resolver problemas práticos? Antes de estudarmos definições formais, veremos como conceitos intuitivos de causalidade podem ser necessários para resolver questões usuais em Inferência Estatística. Para tal, a seguir estudaremos um exemplo de Glymour et al. (2016).

1.1. O Paradoxo de Simpson

Considere que observamos em 500 pacientes 3 variáveis: T e C são as indicadoras de que, respectivamente, o paciente recebeu um tratamento e o paciente curou de uma doença, e Z é uma variável binária cujo significado será discutido mais tarde. Os dados foram resumidos na tabela 1.1.

Em uma primeira análise desta tabela, podemos verificar a efetividade do tratamento dentro de cada valor de Z. Por exemplo, quando $Z = 0$, a frequência de recuperação dentre aqueles que receberam e não receberam o tratamento são, respectivamente: $\frac{81}{6+81} \approx 0.93$ e $\frac{234}{36+234} \approx 0.87$. Similarmente, quando $Z = 1$, as respectivas frequências são: $\frac{192}{71+192} \approx 0.73$ e $\frac{55}{25+55} \approx 0.69$. À primeira vista, para todos os valores de Z, a taxa de recuperação é maior com o tratamento do que sem ele. Isso nos traz informação de que o tratamento é efetivo na recuperação do paciente?

Em uma segunda análise, podemos considerar apenas as contagens para as variáveis T e C, sem estratificar por Z. Dentre os pacientes que receberam e não receberam o tratamento as taxas de recuperação são, respectivamente: $\frac{81+192}{6+71+81+192} \approx 0.78$ e $\frac{234+55}{36+25+234+55} \approx 0.83$. Isto é, sem estratificar por Z, a frequência de recuperação é maior dentre aqueles que não receberam o tratamento do que dentre aqueles que o receberam.

O que é possível concluir destas análises? Uma conclusão ingênua poderia ser a de que, se Z não for observada, então o tratamento não é recomendado. Por outro lado, se Z é observada, não importa qual seja o seu valor, o tratamento será recomendado. A falta de sentido desta conclusão ingênua é o que tornou este tipo de dado famoso como sendo um caso de Paradoxo de Simpson (Simpson, 1951).

Contudo, se a conclusão ingênua é paradoxal e incorreta, então qual conclusão pode ser obtida destes dados? A primeira lição que verificaremos é que não é possível obter uma conclusão sobre o **efeito causal** do tratamento usando apenas a informação na tabela, isto é, associações. Para tal, analisaremos a tabela dando dois nomes distintos para a variável Z. Veremos que, usando exatamente os mesmos dados, uma conclusão válida diferente

```
##      C   0   1
## Z T
## 0 0     36 234
##   1      6  81
## 1 0     25  55
##   1     71 192
```

Tabela 1.1.: Tabela de frequência conjunta das variáveis binárias T, C, e Z.

é obtida para cada nome de Z. Em outras palavras, o efeito causal depende de mais informação do que somente aquela disponível na tabela.

Em um primeiro cenário, considere que Z é a indicadora de que o sexo do paciente é masculino. Observando a tabela, notamos que, proporcionalmente, mais homens receberam o tratamento do que mulheres. Como o tratamento não tem qualquer influência sobre o sexo do paciente, podemos imaginar um cenário em que, proporcionalmente, mais homens escolheram receber o tratamento do que mulheres.

Usando esta observação, podemos fazer sentido do Paradoxo anteriormente obtido. Quando agregamos os dados, notamos que o primeiro grupo de pacientes que receberam o tratamento é predominantemente composto por homens e, similarmente, o segundo grupo de pacientes que não receberam o tratamento é predominantemente composto por mulheres. Isto é, na análise dos dados agregados estamos essencialmente comparando a taxa de recuperação de homens que receberam o tratamento com a de mulheres que não receberam o tratamento. Se assumirmos que, independentemente do tratamento, mulheres tem uma probabilidade de recuperação maior do que homens, então a taxa de recuperação menor no primeiro grupo pode ser explicada pelo fato de ele ser composto predominantemente por homens e não pelo fato de ser o grupo de pacientes que recebeu o tratamento. Também, da análise anterior, obtemos que para cada sexo, a taxa de recuperação é maior com o tratamento do que sem ele. Isto é, neste cenário, o tratamento parece efetivo para a recuperação dos pacientes. Isto significa que a análise estratificando Z é sempre a correta?

Caso o significado da variável Z seja outro, veremos que esta conclusão é incorreta. Considere que Z é a indicadora de que a pressão sanguínea do paciente está elevada. Além disso, é sabido que o tratamento tem como efeito colateral aumentar o risco de pressão elevada nos pacientes. Neste caso, o fato de que há mais indivíduos com pressão elevada dentre aqueles que receberam o tratamento é um efeito direto do tratamento.

Usando esta observação, podemos chegar a outras conclusões sobre o efeito do tratamento sobre a recuperação dos pacientes. Para tal, considere que o tratamento tem um efeito positivo moderado sobre a recuperação dos pacientes, mas que a pressão sanguínea elevada prejudica gravemente a recuperação. Quando fazemos comparações apenas dentre indivíduos com pressão alta ou apenas dentre indivíduos sem pressão alta, não é possível identificar o efeito coletaral do tratamento. Isto é, observamos apenas o efeito positivo moderado que o tratamento tem sobre a recuperação. Por outro lado, quando fazemos a análise agregada, observamos que a frequência de recuperação é maior dentre os indivíduos que não receberam o tratamento do que dentre os que o receberam. Isso ocorre pois o efeito colateral negativo tem um impacto maior sobre a recuperação do paciente do que o efeito geral benéfico. Assim, neste cenário, o tratamento não é eficiente para levar à recuperação do paciente.

Como nossas conclusões dependem de qual história adotamos, podemos ver que a mera apresentação da tabela é insuficiente para determinar a eficiência do tratamento. Observando com cuidado os cenários, identificamos uma explicação geral para as diferentes conclusões. No primeiro cenário, quando Z é sexo, Z é uma causa do indivíduo receber ou não o tratamento. Já no segundo cenário, quando Z é pressão elevada, o tratamento é causa de Z. Isto é, a diferença nas relações entre as variáveis explica as diferenças entre as conclusões obtidas.

Ao longo do curso, desenvolveremos ferramentas para formalizar a diferença entre estes cenários e, com base nisso, conseguir estimar o efeito causal que uma variável X tem sobre outra variável Y. Contudo, para tal, será necessário desenvolver um modelo em que seja possível descrever relações causais. Esta questão será tratada no capítulo 2.

1.1.1. Exercícios

Exercício 1.1 (Glymour et al. (2016)[p.6]). Há evidência de que há correlação positiva entre uma pessoa estar atrasada e estar apressada. Isso significa que uma pessoa pode evitar atrasos se não tiver pressa? Justifique sua resposta em palavras.

2. Modelo Estrutural Causal (SCM)

No capítulo 1 vimos que as relações causais entre variáveis são essenciais para conseguirmos determinar o efeito que uma variável pode ter em outra. Contudo, como podemos especificar relações causais formalmente?

Como resposta a esta pergunta iremos definir o Modelo Estrutural Causal (SCM), que permite especificar formalmente relações causais. Para tal, será necessário primeiro introduzir modelos probabilísticos em grafos. Um curso completo sobre estes modelos pode ser encontrado, por exemplo, em Mauá (2022). A seguir, estudaremos resultados essenciais destes modelos.

2.1. Elementos de Modelos Probabilísticos em Grafos

2.1.1. Grafo Direcionado

Definição 2.1. Um **grafo direcionado**, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, é composto por um conjunto de vértices, $\mathcal{V} = \{V_1, \ldots, V_n\}$, e e um conjunto de arestas, $\mathcal{E} = \{E_1, \ldots, E_m\}$, onde cada aresta é um par ordenado de vértices, isto é, $E_i \in \mathcal{V}^2$.

Para auxiliar nossa intuição sobre a Definição 2.1, é comum representarmos o grafo por meio de uma figura. Nesta, representamos cada vértice por meio de um ponto. Além disso, para cada aresta, (V_i, V_j), traçamos uma seta que aponta de V_i para V_j.

Por exemplo, considere que os vértices são $\mathcal{V} = \{V_1, V_2, V_3\}$ e as arestas são $\mathcal{E} = \{(V_1, V_2), (V_1, V_3), (V_2, V_3)\}$. Neste caso, teremos os 3 pontos como vértices e, além disso, traçaremos setas de V_1 para V_2 e para V_3 e, também, de V_2 para V_3. Podemos desenhar este grafo utilizando os pacotes *dagitty* e *ggdag* (Barrett, 2022, Textor et al., 2016):

```r
library(dagitty)
library(ggdag)
library(ggplot2)

# Especificar o grafo
grafo <- dagitty("dag {
    V1 -> { V2 V3 }
    V2 -> V3
}")

# Exibir a figura do grafo
ggdag(grafo, layout = "circle") +
  theme(axis.text.x = element_blank(),
        axis.ticks.x = element_blank(),
        axis.text.y = element_blank(),
```

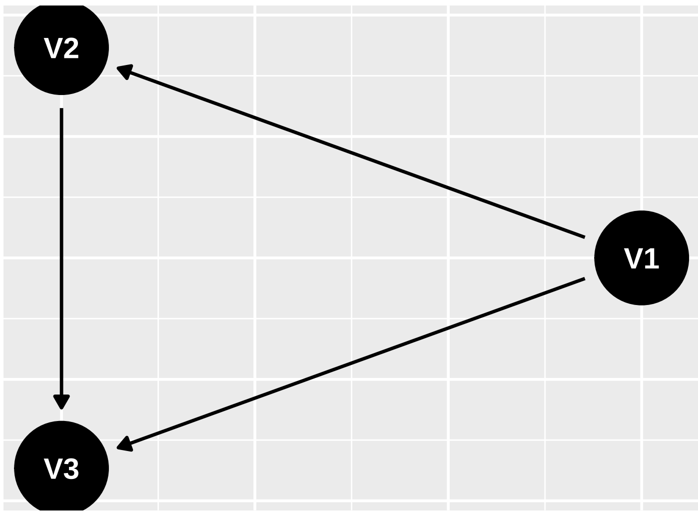

Figura 2.1.: Exemplo de grafo.

```
      axis.ticks.y = element_blank()) +
  xlab("") + ylab("")
```

Grafos direcionados serão úteis para representar causalidade pois seus vértices serão variáveis e suas arestas irão apontar de cada causa imediata para seu efeito. Por exemplo, no Capítulo 1 consideramos um caso em que Sexo e Tratamento são causas imediatas de recuperação e, além disso, Sexo é causa imediata de Tratamento. O grafo na fig. 2.1 poderia representar estas relações se definirmos que V_1 é Sexo, V_2 é Tratamento e V_3 é Recuperação.

Usando a representação de um grafo, podemos imaginar caminhos sobre ele. Um **caminho direcionado** inicia-se em um determinado vértice e, seguindo a direção das setas, vai de um vértice para outro. Por exemplo, (V_1, V_2, V_3) é um caminho direcionado na fig. 2.1, pois existe uma seta de V_1 para V_2 e de V_2 para V_3. É comum denotarmos este caminho direcionado por $V_1 \rightarrow V_2 \rightarrow V_3$. Similarmente, (V_1, V_3, V_2) não é um caminho direcionado, pois não existe seta de V_3 para V_2. A definição de caminho direcionado é formalizada a seguir:

Definição 2.2. Um **caminho direcionado** é uma sequência de vértices em um grafo direcionado, $C = \{C_1, \ldots, C_n\}$ tal que, para cada $1 \leq i < n$, $(C_i, C_{i+1}) \in \mathcal{E}$.

Definição 2.3. Dizemos que V_2 é descendente de V_1 se existe um caminho direcionado de V_1 em V_2.

Um *caminho* é uma generalização de caminho direcionado. Em um caminho, começamos em um vértice e, seguindo por setas, mas não necessariamente na direção em que elas apontam, vamos de um vértice para outro. Por exemplo, na fig. 2.1 vimos que (V_1, V_3, V_2) não é um caminho direcionado pois não existe seta de V_3 para V_2. Contudo, (V_1, V_3, V_2) é um caminho pois existe uma seta ligando V_3 e V_2, a seta que aponta de V_2 para V_3. É comum representarmos este caminho por $V_1 \rightarrow V_3 \leftarrow V_2$. Caminho é formalizado a seguir:

Definição 2.4. Um **caminho** é uma sequência de vértices, $C = \{C_1, \ldots, C_n\}$ tal que, para cada $1 \leq i < n$, $(C_i, C_{i+1}) \in \mathcal{E}$ ou $(C_{i+1}, C_i) \in \mathcal{E}$. $(V_i, V_{i+1}) \in \mathcal{E}$ ou $(V_{i+1}, V_i) \in \mathcal{E}$.

2.1.2. Grafo Direcionado Acíclico (DAG)

Um DAG é um grafo direcionado tal que, para todo vértice, V, não é possível seguir setas partindo de V e voltar para V. Este conceito é formalizado a seguir:

Definição 2.5. Um **grafo direcionado acíclico** (DAG) é um grafo direcionado, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, tal que, para todo vértice, $V \in \mathcal{V}$, não existe um caminho direcionado, $C = \{C_1, \ldots, C_n\}$ tal que $C_1 = V = C_n$.

Usualmente representaremos as relações causais por meio de um DAG. Especificamente, existirá uma aresta de V_1 para V_2 para indicar que V_1 é causa imediata de V_2. Caso um grafo direcionado não seja um DAG, então existe um caminho de V em V, isto é, V seria uma causa de si mesma, o que desejamos evitar.

Um DAG induz uma *ordem parcial* entre os seus vértices. Isto é, se existe uma aresta de V_1 para V_2, então podemos interpretar que V_1 antecede V_2 causalmente. Com base nesta ordem parcial, é possível construir diversas definições que nos serão úteis.

Dizemos que V_1 é pai de V_2 em um DAG, $\mathcal{G}$, se existe uma aresta de V_1 a V_2, isto é, $(V_1, V_2) \in \mathcal{E}$. Denotamos por $Pa(V)$ o conjunto de todos os pais de V:

Definição 2.6. O conjunto de **pais** de $V \in \mathcal{V}$ em um DAG, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, é:

$$Pa(V) := \{V^* \in \mathcal{V} : (V^*, V) \in \mathcal{E}\}.$$

Similarmente, dizemos que V_1 é um ancestral de V_2 em um DAG, se V_1 antecede V_2 causalmente. Isto é, se V_1 é pai de V_2 ou, pai de pai de V_2, ou pai de pai de pai de V_2, e assim por diante ... Denotamos por $Anc(\mathbb{V})$ o conjunto de todos os ancestrais de elementos de $\mathbb{V}$:

Definição 2.7. Em um DAG, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, o conjunto de **ancestrais** de $\mathbb{V} \subseteq \mathcal{V}$, $Anc(\mathbb{V})$, é tal que $Anc(\mathbb{V}) \subseteq \mathcal{V}$ e $V^* \in Anc(\mathbb{V})$ se e somente se existe $V \in \mathbb{V}$ e um caminho direcionado, C, tal que $C_1 = V^*$ e $C_i = V$.

Note que podemos interpretar $Anc(\mathbb{V})$ como o conjunto de todas as causas diretas e indiretas de $\mathbb{V}$.

Finalmente, diremos que um conjunto de vértices, $\mathcal{A} \subseteq \mathcal{V}$ é *ancestral* em um DAG, se não existe algum vértice fora de $\mathcal{A}$ que seja pai de algum vértice em $\mathcal{A}$. Segundo nossa interpretação causal, $\mathcal{A}$ será ancestral quando nenhum vértice fora de $\mathcal{A}$ é causa direta de algum vértice em $\mathcal{A}$:

Definição 2.8. Dizemos que $\mathcal{A} \subseteq \mathcal{V}$ é **ancestral** em um DAG se, para todo vértice $V \in \mathcal{A}$, temos que $Pa(V) \subseteq \mathcal{A}$.

Lema 2.9. *Em um DAG, $\mathcal{G}$, para todo $\mathbb{V} \subseteq \mathcal{V}$, $Anc(\mathbb{V})$ é ancestral.*

2.1.3. Modelo Probabilístico em um DAG

Um modelo probabilístico em um DAG é tal que cada um dos vértices é uma variável aleatória. O DAG será usado para descrever relações de independência condicional existentes entre estas variáveis. Mais especificamente, cada vértice será independente dos demais vértices dados os seus pais. Uma maneira alternativa de pensar sobre esta afirmação é imaginar que cada vértice é gerado somente pelos seus pais. Esta intuição é formalizada em Definição 2.10:

Definição 2.10. Para $\mathcal{V}$ um conjunto de variáveis aleatórias, dizemos que uma função de densidade sobre $\mathcal{V}$, f, é compatível com um DAG, $\mathcal{G}$, se:

$$f(v_1, \ldots, v_n) = \prod_{i=1}^{n} f(v_i | Pa(v_i))$$

Na prática, pode ser difícil verificar se a Definição 2.10 está satisfeita Para esses casos, pode ser útil aplicar o Lema 2.11:

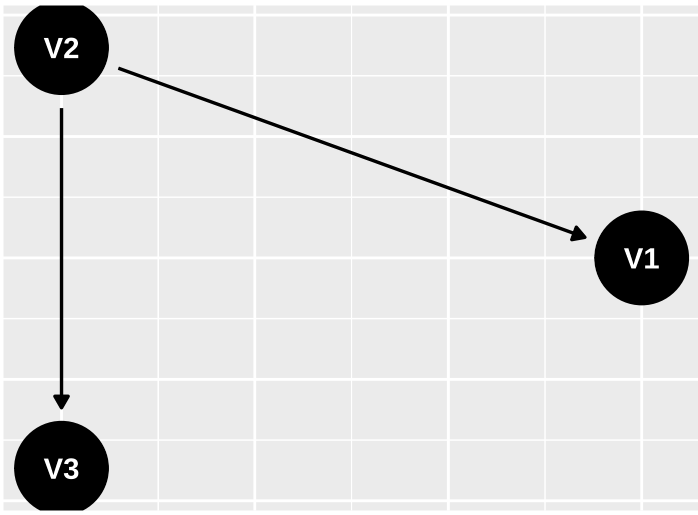

Figura 2.2.: Ilustração de confundidor.

Lema 2.11. *Uma função de densidade, f, é compatível com um DAG, $\mathcal{G}$, se e somente se, existem funções, $g_1, \ldots, g_n$ tais que:*

$$f(v_1, \ldots, v_n) = \prod_{i=1}^{n} g_i(v_i, Pa(v_i))$$

O seguinte lema também é útil

Lema 2.12. *Seja $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ um DAG. Se $\mathcal{A}$ é ancestral e f é compatível com $\mathcal{G}$, então*

$$f(\mathcal{A}) = \prod_{V \in \mathcal{A}} f(V|Pa(V))$$

A seguir, estudaremos três tipos fundamentais de modelos probabilísticos em DAG's com 3 vértices. A intuição obtida a partir destes exemplos continuará valendo quando estudarmos grafos mais gerais.

2.1.4. Exemplos de Modelo Probabilístico em um DAG

Nos exemplos a seguir, considere que $\mathcal{V} = (V_1, V_2, V_3)$.

Confundidor (Confounder)

No modelo de confundidor, as únicas duas arestas são (V_2, V_1) e (V_2, V_3). Uma ilustração de um confundidor pode ser encontrada na fig. 2.2. O modelo de confundidor pode ser usado quando acreditamos que V_2 é uma causa comum a V_1 e a V_3. Além disso, V_1 não é causa imediata de V_3 nem vice-versa.

Em um modelo de confundidor a relação de dependência entre V_1 e V_3 é explicada pelos resultados a seguir:

Lema 2.13. *Para qualquer probabilidade compatível com o DAG na fig. 2.2, $V_1 \perp\!\!\!\perp V_3 | V_2$.*

Demonstração.

$$\begin{aligned} f(v_1, v_3|v_2) &= \frac{f(v_1, v_2, v_3)}{f(v_2)} \\ &= \frac{f(v_2)f(v_1|v_2)f(v_3|v_2)}{f(v_2)} && \text{Definição 2.10} \\ &= f(v_1|v_2)f(v_3|v_2) \end{aligned}$$

□

Lema 2.14. *Existe ao menos uma probabilidade compatível com o DAG na fig. 2.2 tal que* $V_1 \not\perp V_3$.

Demonstração. Considere que $V_2 \sim \text{Bernoulli}(0.02)$. Além disso, $V_1, V_3 \in \{0,1\}$ são independentes dado V_2. Também, $\mathbb{P}(V_1 = 1|V_2 = 1) = \mathbb{P}(V_3 = 1|V_2 = 1) = 0.9$ e $\mathbb{P}(V_1 = 1|V_2 = 0) = \mathbb{P}(V_3 = 1|V_2 = 0) = 0.05$. Note que, por construção, $\mathbb{P}$ é compatível com fig. 2.2. Isto é, $P(v_1, v_2, v_3) = \mathbb{P}(v_2)\mathbb{P}(v_1|v_2)\mathbb{P}(v_3|v_2)$. Além disso,

$$\begin{aligned} \mathbb{P}(V_1 = 1) &= \mathbb{P}(V_1 = 1, V_2 = 1) + \mathbb{P}(V_1 = 1, V_2 = 0) \\ &= \mathbb{P}(V_2 = 1)\mathbb{P}(V_1 = 1|V_2 = 1) + \mathbb{P}(V_2 = 0)\mathbb{P}(V_1 = 1|V_2 = 0) \\ &= 0.02 \cdot 0.9 + 0.98 \cdot 0.05 = 0.067 \end{aligned}$$

Por simetria, $\mathbb{P}(V_3 = 1) = 0.067$. Além disso,

$$\begin{aligned} \mathbb{P}(V_1 = 1, V_3 = 1) &= \mathbb{P}(V_1 = 1, V_3 = 1, V_2 = 1) + \mathbb{P}(V_1 = 1, V_3 = 1, V_2 = 0) \\ &= \mathbb{P}(V_2 = 1)\mathbb{P}(V_1 = 1|V_2 = 1)\mathbb{P}(V_3 = 1|V_2 = 1) + \mathbb{P}(V_2 = 0)\mathbb{P}(V_1 = 1|V_2 = 0)\mathbb{P}(V_3 = 1|V_2 = 0) \\ &= 0.02 \cdot 0.9 \cdot 0.9 + 0.98 \cdot 0.05 \cdot 0.05 = 0.01865 \end{aligned}$$

Como $\mathbb{P}(V_1 = 1)\mathbb{P}(V_3 = 1) = 0.067 \cdot 0.067 \approx 0.0045 \neq 0.01865 = \mathbb{P}(V_1 = 1, V_3 = 1)$, temos que V_1 e V_3 não são independentes. □

Combinando os Lemas 2.13 e 2.14 é possível compreender melhor como usaremos confundidores num contexto causal. Nestes casos, V_2 será uma causa comum a V_1 e a V_3. Esta causa comum torna V_1 e V_3 associados, ainda que nenhum seja causa direta ou indireta do outro.

Podemos contextualizar estas ideias em um caso de diagnóstico de dengue. Considere que V_2 é a indicadora de que um indivíduo tem dengue, e V_1 e V_3 são indicadoras de sintomas típicos de dengue, como dor atrás dos olhos e febre. Neste caso, V_1 e V_3 tipicamente são associados: caso um paciente tenha febre, aumenta a probabilidade de que tenha dengue e, portanto, aumenta a probabilidade de que tenha dor atrás dos olhos. Contudo, apesar dessa associação V_3 não tem influência causal sobre V_1. Se aumentarmos a temperatura corporal do indivíduo, não aumentará a probabilidade de que ele tenha dor atrás dos olhos. A dengue que causa febre, não o contrário.

Cadeia (Chain)

No modelo de cadeia, as únicas duas arestas são (V_1, V_2) e (V_2, V_3). Uma ilustração de uma cadeia pode ser encontrada na fig. 2.3. Neste modelo, acreditamos que V_1 é causa de V_2 que, por sua vez, é causa de V_3. Assim, V_1 é ancestral de V_3, isto é, o primeiro é causa indireta do segundo.

Em um modelo de cadeia a relação de dependência entre V_1 e V_3 é explicada pelos resultados a seguir:

Lema 2.15. *Para qualquer probabilidade compatível com o DAG na fig. 2.3,* $V_1 \perp V_3|V_2$.

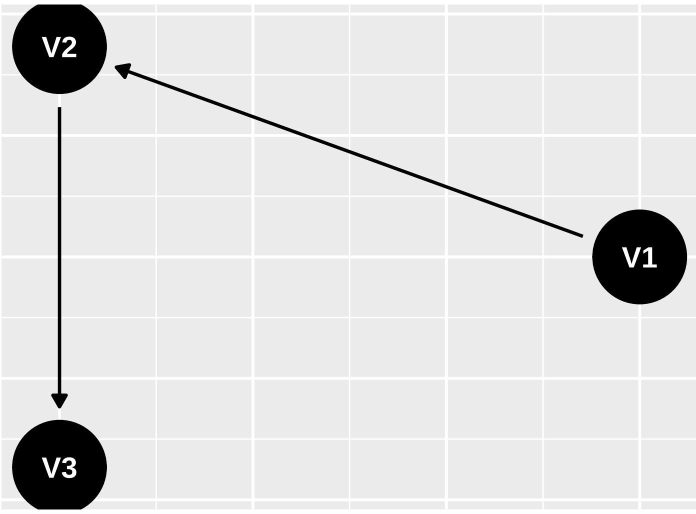

Figura 2.3.: Ilustração de cadeia.

Demonstração.

$$\begin{aligned} f(v_3|v_1,v_2) &= \frac{f(v_1,v_2,v_3)}{f(v_1,v_2)} \\ &= \frac{f(v_1)f(v_2|v_1)f(v_3|v_2)}{f(v_1)f(v_2|v_1)} && \text{Definição 2.10} \\ &= f(v_3|v_2) \end{aligned}$$

□

Lema 2.16. *Existe ao menos uma probabilidade compatível com o DAG na fig. 2.3 tal que* $V_1 \not\perp V_3$.

Demonstração. Considere que $V_1 \sim \text{Bernoulli}(0.5)$, $\mathbb{P}(V_2 = 1|V_1 = 1) = 0.9$, $\mathbb{P}(V_2 = 1|V_1 = 0) = 0.05$, $\mathbb{P}(V_3 = 1|V_2 = 1, V_1) = 0.9$, e $\mathbb{P}(V_3 = 1|V_2 = 0, V_1) = 0.05$. Note que (V_1, V_2, V_3) formam uma Cadeia de Markov. Note que, por construção, $\mathbb{P}$ é compatível com fig. 2.3. Isto é, $P(v_1, v_2, v_3) = \mathbb{P}(v_1)\mathbb{P}(v_2|v_1)\mathbb{P}(v_3|v_2)$. Além disso,

$$\begin{aligned} \mathbb{P}(V_3 = 1) &= \mathbb{P}(V_1 = 0, V_2 = 0, V_3 = 1) + \mathbb{P}(V_1 = 0, V_2 = 1, V_3 = 1) \\ &+ \mathbb{P}(V_1 = 1, V_2 = 0, V_3 = 1) + \mathbb{P}(V_1 = 1, V_2 = 1, V_3 = 1) \\ &= 0.5 \cdot 0.9 \cdot 0.05 + 0.5 \cdot 0.05 \cdot 0.9 \\ &+ 0.5 \cdot 0.05 \cdot 0.05 + 0.5 \cdot 0.9 \cdot 0.9 = 0.45125 \end{aligned}$$

Além disso,

$$\begin{aligned} \mathbb{P}(V_1 = 1, V_3 = 1) &= \mathbb{P}(V_1 = 1, V_2 = 0, V_3 = 1) + \mathbb{P}(V_1 = 1, V_2 = 1, V_3 = 1) \\ &= 0.5 \cdot 0.05 \cdot 0.9 + 0.5 \cdot 0.9 \cdot 0.9 = 0.40625 \end{aligned}$$

Como $\mathbb{P}(V_1 = 1)\mathbb{P}(V_3 = 1) = 0.5 \cdot 0.45125 \approx 0.226 \neq 0.40625 = \mathbb{P}(V_1 = 1, V_3 = 1)$, temos que V_1 e V_3 não são independentes. □

Combinando os Lemas 2.15 e 2.16 é possível compreender melhor como usaremos cadeias num contexto causal.

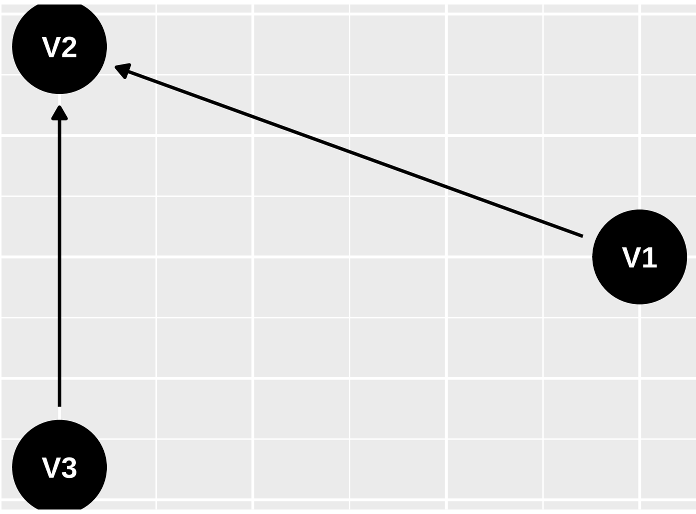

Figura 2.4.: Ilustração de colisor.

Nestes casos, V_2 será uma consequência de V_1 e uma causa de V_3. Assim, a cadeia torna V_1 e V_3 e associados, ainda que nenhum seja causa direta do outro. Contudo, ao contrário do confundidor, neste caso V_1 é uma causa indireta de V_3, isto é, tem influência causal sobre V_3.

Para contextualizar estas ideias, considere que V_1 é a indicadora de consumo elevado de sal, V_2 é a indicadora de pressão alta, e V_3 é a indicadora de ocorrência de um derrame. Como consumo elevado de sal causa pressão alta e pressão alta tem influência causal sobre a ocorrência de um derrame, pressão alta é uma cadeia que é um mediador entre consumo elevado de sal e ocorrência de derrame. Assim, consumo elevado de sal tem influência causal sobre a ocorrência de derrame.

Colisor (Collider)

O último exemplo de DAG com 3 vértices que estudaremos é o de modelo de colisor, em que as únicas duas arestas são (V_1, V_2) e (V_3, V_2). Uma ilustração de um colisor pode ser encontrada na fig. 2.4. O modelo de colisor pode ser usado quando acreditamos que V_1 e V_3 são causas comuns a V_2. Além disso, V_1 não é causa imediata de V_3 nem vice-versa.

Em um modelo de colisor a relação de dependência entre V_1 e V_3 é explicada pelos resultados a seguir:

Lema 2.17. *Para qualquer probabilidade compatível com o DAG na fig. 2.4, $V_1 \perp\!\!\!\perp V_3$.*

Demonstração.

$$\begin{aligned}
f(v_1, v_3) &= \int f(v_1, v_2, v_3)dv_2 \\
&= \int f(v_1)f(v_3)f(v_2|v_1, v_3)dv_2 && \text{Definição 2.10} \\
&= f(v_1)f(v_3)\int f(v_2|v_1, v_3)dv_2 \\
&= f(v_1)f(v_3)
\end{aligned}$$

□

Lema 2.18. *Existe ao menos uma probabilidade compatível com o DAG na fig. 2.4 tal que* $V_1 \not\perp V_3 | V_2$.

Demonstração. Considere que V_1 e V_3 são independentes e tem distribuição Bernoulli(0.5). Além disso, $V_2 \equiv V_1 + V_3$. Como $\mathbb{P}(V_3 = 1) = 0.5$ e $\mathbb{P}(V_3 = 1 | V_1 = 1, V_2 = 2) = 1$, conclua que $V_1 \not\perp V_3 | V_2$. □

Combinando os Lemas 2.17 e 2.18 vemos como utilizaremos confundidores num contexto causal. Nestes casos, V_1 e V_3 serão causas comuns e independentes de V_2. Uma vez que obtemos informação sobre o efeito comum, V_2, V_1 e V_3 passam a ser associados.

Esse modelo pode ser contextualizado observando a prevalência de doenças em uma determinada população (Sackett, 1979). Considere que V_1 e V_3 são indicadoras de que um indivíduo tem doenças que ocorrem independentemente na população. Além disso, V_2 é a indicadora de que o indíviduo foi hospitalizado, isto é, V_2 é influeciado causalmente tanto por V_1 quanto por V_3. Para facilitar as contas envolvidas, desenvolveremos o exemplo com distribuições fictícias. Considere que V_1 e V_3 são independentes e tem distribuição Bernoulli(0.05). Além disso, quanto maior o número de doenças, maior a probabilidade de o indíviduo ser hospitalizado. Por exemplo, $\mathbb{P}(V_2 = 1 | V_1 = 0, V_3 = 0) = 0.01$, $\mathbb{P}(V_2 = 1 | V_1 = 0, V_3 = 1) = 0.1$, $\mathbb{P}(V_2 = 1 | V_1 = 1, V_3 = 0) = 0.1$, e $\mathbb{P}(V_2 = 1 | V_1 = 1, V_3 = 1) = 0.5$.

Com base nestas especificações, podemos verificar se V_1 e V_3 estão associados quando $V_2 = 1$. Para tal, primeiramente calcularemos algumas probabilidades conjuntas que serão úteis:

$$\begin{cases} \mathbb{P}(V_1 = 0, V_2 = 1, V_3 = 0) & = 0.95 \cdot 0.01 \cdot 0.95 = 0.009025 \\ \mathbb{P}(V_1 = 0, V_2 = 1, V_3 = 1) & = 0.95 \cdot 0.1 \cdot 0.05 = 0.0475 \\ \mathbb{P}(V_1 = 1, V_2 = 1, V_3 = 0) & = 0.05 \cdot 0.1 \cdot 0.95 = 0.0475 \\ \mathbb{P}(V_1 = 1, V_2 = 1, V_3 = 1) & = 0.05 \cdot 0.5 \cdot 0.05 = 0.00125 \end{cases} \tag{2.1}$$

Com base nestes cálculos é possível obter a prevalência da doença dentre os indivíduos hospitalizados:

$$\begin{aligned} \mathbb{P}(V_1 = 1 | V_2 = 1) &= \frac{\mathbb{P}(V_1 = 1, V_2 = 1)}{\mathbb{P}(V_2 = 1)} \\ &= \frac{0.0475 + 0.00125}{0.009025 + 0.0475 + 0.0475 + 0.00125} && \text{eq. (2.1)} \\ &\approx 0.46 \end{aligned}$$

Finalmente,

$$\begin{aligned} \mathbb{P}(V_1 = 1 | V_2 = 1, V_3 = 1) &= \frac{\mathbb{P}(V_1 = 1, V_2 = 1, V_3 = 1)}{\mathbb{P}(V_2 = 1, V_3 = 1)} \\ &= \frac{\mathbb{P}(V_1 = 1, V_2 = 1, V_3 = 1)}{\mathbb{P}(V_1 = 0, V_2 = 1, V_3 = 1) + \mathbb{P}(V_1 = 1, V_2 = 1, V_3 = 1)} \\ &= \frac{0.00125}{0.0475 + 0.00125} && \text{eq. (2.1)} \\ &\approx 0.26 \end{aligned}$$

Como $\mathbb{P}(V_1 = 1 | V_2 = 1) = 0.46 \neq 0.26 \approx \mathbb{P}(V_1 = 1 | V_2 = 1, V_3 = 1)$, verificamos que V_1 não é independente de V_3 dado V_2. De fato, ao observar que um indivíduo está hospitalizado e tem uma das doenças, a probabilidade de que ele tenha a outra doença é inferior àquela obtida se soubéssemos apenas que o indivíduo está hospitalizado.

Esta observação não implica que uma doença tenha influência causal sobre a outra. Note que a frequência de hospitalização aumenta drasticamente quando um indivíduo tem ao menos uma das doenças. Além disso, cada uma das doenças é relativamente rara na população geral. Assim, dentre os indíviduos hospitalizados, a frequência daqueles que tem somente uma das doenças é maior do que seria caso as doenças não estivessem associadas. Quando fixamos o valor de uma consequência comum (hospitalização), as causas (doenças) passam a ser associadas. Esta associação não significa que infectar um indivíduo com uma das doenças reduz a probabilidade que ele tenha a outra.

2.1.5. Modelo Estrutural Causal (Structural Causal Model)

Com base nos conceitos abordados anteriormente, finalmente podemos definir formalmente o Modelo Estrutural Causal (SCM):

Definição 2.19. Um SCM é um par $(\mathcal{G}, f)$ tal que $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ é um DAG (Definição 2.5) e f é uma função de densidade sobre $\mathcal{V}$ compatível com $\mathcal{G}$ (Definição 2.10). Neste caso, é comum chamarmos $\mathcal{G}$ de **grafo causal** do SCM $(\mathcal{G}, f)$.

Note pela Definição 2.19 que um SCM é formalmente um modelo probabilístico em um DAG. O principal atributo de um SCM que o diferencia de um modelo probabilístico genérico em um DAG é como o interpretamos. Existe uma aresta de V_1 em V_2 em um SCM se e somente se V_1 é uma causa direta de V_2.

Definição 2.20. Dizemos que $(\mathcal{G}, f)$ é um SCM linear Gaussiano se, existe matriz diagonal positiva, Σ, e $\beta \in \Re^{n \times n}$ tal que, para todo vértice V_i, $\beta_{i,j} = 0$ quando $V_j \notin Pa(V_i)$ e

$$V_i|Pa(V) \sim N(\beta_i, \Sigma_{i,i})$$

Lema 2.21. *Se $(\mathcal{G}, f)$ é um SCM linear Gaussiano, então $\mathcal{V} \sim N(0, (I - \beta)^{-1}\Sigma((I - \beta)^{-1})^t)$.*

No próximo capítulo estudaremos consequências desta interpretação causal. Contudo, antes disso, a próxima seção desenvolverá um resultado fundamental de modelos probabilísticos em DAGs que será fundamental nos capítulos posteriores.

2.1.6. Exercícios

Exercício 2.22. Em um DAG, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, Considere que $Anc^*(\mathbb{V}) \subseteq \mathcal{V}$ é definido como o menor conjunto tal que $\mathbb{V} \subseteq Anc^*(\mathbb{V})$ e, se $V \in Anc^*(\mathbb{V})$, então $Pa(V) \subseteq Anc^*(\mathbb{V})$. Prove que $Anc(\mathbb{V}) \equiv Anc^*(\mathbb{V})$.

Exercício 2.23. Prove o Lema 2.9.

Exercício 2.24. Prove que se $\mathbf{Z}$ é ancestral, então $f(\mathbf{Z}) = \prod_{Z \in \mathbf{Z}} f(Z|Pa(Z))$.

Exercício 2.25. Sejam $\mathcal{G}_1 = (\mathcal{V}, \mathcal{E}_1)$ e $\mathcal{G}_2 = (\mathcal{V}, \mathcal{E}_2)$ grafos tais que $\mathcal{E}_1 \subseteq \mathcal{E}_2$. Prove que se f é compatível com $\mathcal{G}_2$, então f é compatível com $\mathcal{G}_1$.

Exercício 2.26. Prove o Lema 2.11.

Exercício 2.27. Prove o Lema 2.12.

Exercício 2.28. Considere que (X_1, X_2) são independentes e tais que $\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = 0.5$. Além disso, $Y \equiv X_1 \cdot X_2$.

(a) Desenhe um DAG compatível com as relações de independência dadas pelo enunciado.

(b) Prove que Y e X_1 são independentes. Isso contradiz sua resposta para o item anterior?

Exercício 2.29. Para cada um dos modelos de confundidor, cadeia e colisor, dê exemplos de situações práticas em que este modelo é razoável.

Exercício 2.30. Considere que, dado T, $X_1, \ldots, X_n$ são i.i.d. e $X_i|T \sim \text{Bernoulli}(T)$. Além disso, $T \sim \text{Beta}(a, b)$.

(a) Seja $f(t, x_1, \ldots, x_n)$ dada pelo enunciado. Exiba um DAG, $\mathcal{G}$, tal que f é compatível com $\mathcal{G}$.

(b) $(X_1, \ldots, X_n)$ são independentes?

(c) Determine $f(x_1, \ldots, x_n)$.

Exercício 2.31. Exiba um exemplo em que V_1, V_2, V_3 sejam binárias, que V_2 seja um colisor e que, além disso, $Corr[V_1, V_3|V_2 = 1] > 0$.

Exercício 2.32. Seja $\mathcal{V} = (V_1, V_2, V_3)$ Exiba um exemplo de f sobre $\mathcal{V}$ e grafos $\mathcal{G}_1$ e $\mathcal{G}_2$ sobre $\mathcal{V}$ tais que $\mathcal{G}_1 \neq \mathcal{G}_2$ e f é compatível tanto com $\mathcal{G}_1$ quanto com $\mathcal{G}_2$.

Exercício 2.33. Seja f uma densidade arbitrária sobre $\mathcal{V} = (V_1, \ldots, V_n)$. Exiba um DAG sobre $\mathcal{V}$, $\mathcal{G}$, tal que f é compatível com $\mathcal{G}$.

Exercício 2.34. Exiba um exemplo em que V_2 é um colisor entre V_1 e V_2, V_4 tem como único pai V_2 e V_1 e V_3 são dependentes dado V_4.

2.2. Independência Condicional e D-separação

Independência condicional é uma forma fundamental de indicar relações entre variáveis aleatórias. Se $\mathbf{X}_1, \ldots, \mathbf{X}_d$ e $\mathbf{Y}$ são vetores de variáveis aleatórias, definimos que $(\mathbf{X}_1, \ldots, \mathbf{X}_d)|\mathbf{Y}$, isto é, $\mathbf{X}_1, \ldots, \mathbf{X}_d$ são independentes dado $\mathbf{Y}$, se conhecido o valor de $\mathbf{Y}$, observar quaisquer valores de $\mathbf{X}$ não traz informação sobre os demais valores. Nesta seção veremos que as relações de independência condicional em um SCM estão diretamente ligadas ao seu grafo.

2.2.1. Independência Condicional

Definição 2.35. Dizemos que $(\mathbf{X}_1, \ldots, \mathbf{X}_d)$ são independentes dado $\mathbf{Y}$ se, para qualquer $\mathbf{x}_1, \ldots, \mathbf{x}_n$ e $\mathbf{y}$,

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_n|\mathbf{y}) = \prod_{i=1}^{d} f(\mathbf{x}_i|\mathbf{y})$$

Em particular, $(\mathbf{X}_1, \ldots, \mathbf{X}_d)$ são independentes se, para quaiquer $(\mathbf{x}_1, \ldots, \mathbf{x}_d)$,

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_d) = \prod_{i=1}^{d} f(\mathbf{x}_i)$$

Verificar se a Definição 2.35 está satisfeita nem sempre é fácil. A princípio, ela exige obter tanto a distribuição condicional conjunta, $f(\mathbf{x}_1, \ldots, \mathbf{x}_d|\mathbf{y})$, quanto cada uma das marginais, $f(\mathbf{x}_i|\mathbf{y})$. O Lema 2.36 a seguir apresenta outras condições que são equivalentes a independência condicional:

Lema 2.36. *As seguintes afirmações são equivalentes:*

1. $(\mathbf{X}_1, \ldots, \mathbf{X}_d)$ *são independentes dado* $\mathbf{Y}$*,*

2. *Existem funções,* $h_1, \ldots, h_d$ *tais que* $f(\mathbf{x}_1, \ldots, \mathbf{x}_d|\mathbf{y}) = \prod_{j=1}^{d} h_j(\mathbf{x}_j, \mathbf{y})$*.*

3. *Para todo* i*,* $f(\mathbf{x}_i|\mathbf{x}_{-i}, \mathbf{y}) = f(\mathbf{x}_i|\mathbf{y})$*.*

4. *Para todo* i*,* $f(\mathbf{x}_i|\mathbf{x}_1^{i-1}, \mathbf{y}) = f(\mathbf{x}_i|\mathbf{y})$*.*

As condições no Lema 2.36 são, em geral, mais fáceis de verificar do que a definição direta de independência condicional. A seguir veremos que, em um SMC, pode ser mais fácil ainda verificar muitas das relações de independência condicional.

2.2.2. D-separação

Em um SCM, é possível indicar as relações de independência incondicional em $\mathcal{V}$ por meio do grafo associado. Intuitivamente, haverá uma dependência entre V_1 e V_2 se for possível transmitir a informação de V_1 para V_2 por um caminho que ligue ambos os vértices. Para entender se a informação pode ser transmitida por um caminho, classificaremos a seguir os vértices que o constituem.

Definição 2.37. Seja $C = (C_1, \ldots, C_n)$ um caminho em um DAG, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Para cada $2 \leq i \leq n-1$:

- C_i é um **colisor** em C se $(C_{i-1}, C_i) \in \mathcal{E}$ e $(C_{i+1}, C_i) \in \mathcal{E}$, isto é, existem arestas apontando de C_{i-1} e de C_{i+1} para C_i. Neste caso, desenhamos $C_{i-1} \rightarrow C_i \leftarrow C_{i+1}$.

Note que a classificação na Definição 2.37 generaliza os exemplos de DAG's com 3 vértices na Seção 2.1.4.

Essa classificação é ilustrada com o DAG na fig. 2.5. Existem dois caminhos que vão de V_1 a V_4: $V_1 \rightarrow V_2 \leftarrow V_4$ e $V_1 \rightarrow V_2 \rightarrow V_3 \leftarrow V_4$. No primeiro caminho V_2 é um colisor, pois o caminho passa por duas arestas que apontam para V_2. Já no segundo caminho V_2 é uma cadeia e V_3 é um colisor. Note que a classificação do vértice depende do caminho analisado. Enquanto que no primeiro caminho V_2 é um colisor, no segundo V_2 é uma cadeia

Com base na Definição 2.37, é possível compreender se um caminho permite a passagem de informação. Na Seção 2.1.4 vimos que, se V_2 não é um colisor entre V_1 e V_3, então V_1 e V_3 são independentes dado V_2. Por analogia, podemos intuir que um vértice que não é um colisor num caminho não permite a passagem de informação quando seu valor é conhecido. Similarmente, na Seção 2.1.4, se V_2 é um colisor entre V_1 e V_3, então V_1 e V_3 são independentes. Assim, também podemos intuir que um vértice que é um colisor em um caminho não permite a passagem de informação quando seu valor é desconhecido. Finalmente, a informação não passa pelo caminho quando ela não passa por pelo menos um de seus vértices. Neste caso, dizemos que o caminho está *bloqueado*:

Definição 2.38. Seja $C = (C_1, \ldots, C_n)$ um caminho em um DAG, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Dizemos que C está bloqueado dado $\mathbf{Z} \subset \mathcal{V}$, se

1. Existe algum $2 \leq i \leq n-1$ tal que C_i não é um colisor em C e $C_i \in \mathbf{Z}$, ou

2. Existe algum $2 \leq i \leq n-1$ tal que C_i é um colisor em C e $C_i \notin Anc(\mathbf{Z})$.

Finalmente, dizemos que $\mathbb{V}_1$ está d-separado de $\mathbb{V}_2$ dado $\mathbb{V}_3$ se todos os caminhos de $\mathbb{V}_1$ a $\mathbb{V}_2$ estão bloqueados dado $\mathbb{V}_3$:

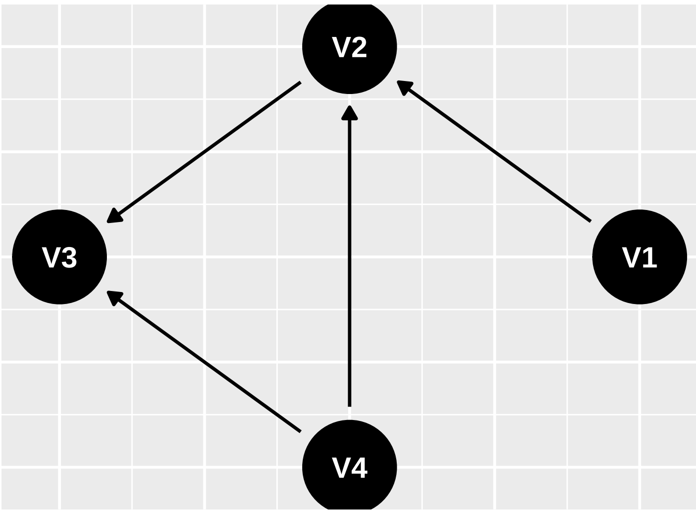

Figura 2.5.: Ilustração do conceito de bloqueio de um caminho. No caminho (V1, V2, V4), V2 é um colisor. Isto ocorre pois, para chegar de V1 a V4 passando apenas por V2, as duas arestas apontam para V2. Já no caminho (V1, V2, V3, V4) temos que V2 é uma cadeia. Para chegar de V1 a V3 passando por V2, passa-se por duas arestas, uma entrando e outra saindo de V2. Como V2 é um colisor em (V1, V2, V4), este caminho está bloqueado se e somente se o valor de V2 é desconhecido. Como V2 é uma cadeia em (V1, V2, V3, V4), esse caminho está bloqueado quando o valor de V2 é conhecido.

Definição 2.39. Seja $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ um DAG. Para $\mathbb{V}_1, \mathbb{V}_2, \mathbb{V}_3 \subseteq \mathcal{V}$, dizemos que $\mathbb{V}_1$ está d-separado de $\mathbb{V}_2$ dado $\mathbb{V}_3$ se, para todo caminho $C = (C_1, \ldots, C_n)$ tal que $C_1 \in V_1$ e $C_n \in \mathbb{V}_2$, C está bloqueado dado $\mathbb{V}_3$. Neste caso, escrevemos $\mathbb{V}_1 \perp \mathbb{V}_2 | \mathbb{V}_3$.

Intuitivamente, se $\mathbb{V}_1 \perp \mathbb{V}_2 | \mathbb{V}_3$, então não é possível passar informação de $\mathbb{V}_1$ a $\mathbb{V}_2$ quando $\mathbb{V}_3$ é conhecido. Assim, temos razão para acreditar que $\mathbb{V}_1$ é condicionalmente independente de $\mathbb{V}_2$ dado $\mathbb{V}_3$, isto é $\mathbb{V}_1 \perp\!\!\!\perp \mathbb{V}_2 | \mathbb{V}_3$. Esta conclusão é apresentada no Teorema 2.40 a seguir:

Teorema 2.40. *Seja $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ um DAG e $\mathcal{V}$ um conjunto de variáveis aleatórias. $\mathbb{V}_1$ está d-separado de $\mathbb{V}_2$ dado $\mathbb{V}_3$ se e somente se, para todo f compatível com $\mathcal{G}$, $\mathbb{V}_1 \perp\!\!\!\perp^f \mathbb{V}_2 | \mathbb{V}_3$.*

Exemplo 2.41. Considere o DAG na fig. 2.5. Para avaliar se V_1 e V_3 são d-separados, precisamos analisar todos os caminhos de um para o outro. Estes caminhos são: $V_1 \rightarrow V_2 \rightarrow V_3$, e $V_1 \rightarrow V_2 \leftarrow V_4 \rightarrow V_3$. No primeiro caminho V_2 não é um colisor e, assim, o caminho não está bloqueado marginalmente. Portanto, V_1 e V_3 não são d-separados marginalmente. Por outro lado, no segundo caminho V_2 é um colisor e V_4 não o é. Assim, condicionando em V_2, este caminho não está bloqueado. Portanto, V_1 e V_3 não são d-separados dado V_2. Finalmente, dado V_2 e V_4, ambos os caminhos estão bloqueados, pois V_2 não é um colisor no primeiro e V_4 não é um colisor no segundo. Assim, V_1 e V_3 são d-separados dado (V_2, V_4). Para treinar este raciocínio, continue analisando a d-separação entre V_1 e V_4.

O algoritmo para testar d-separação está implementado em diversos pacotes. Além disso, é possível utilizar o Teorema 2.40 para enunciar todas as relações de independência condicional que são necessárias em um grafo. Estas implementações estão ilustradas abaixo:

```r
# Especificar o grafo
grafo <- "dag{
   V1 -> V2 <- V4;
   V2 -> V3 <- V4
}"

dseparated(grafo, "V1", "V3", c("V2"))

## [1] FALSE

dseparated(grafo, "V1", "V3", c("V4"))

## [1] FALSE

dseparated(grafo, "V1", "V3", c("V2", "V4"))

## [1] TRUE

impliedConditionalIndependencies(grafo)

## V1 _||_ V3 | V2, V4
## V1 _||_ V4
```

Exemplo 2.42. Considere que $\mathbb{V}_1$ e $\mathbb{V}_2$ não são d-separados dado $\mathbb{V}_3$. O Teorema 2.40 garante apenas que existe algum f compatível com o DAG tal que $\mathbb{V}_1$ e $\mathbb{V}_2$ são condicionalmente dependentes dado $\mathbb{V}_3$ segundo f. É possível mostrar que o conjunto de f's compatíveis com o grafo em que $\mathbb{V}_1$ e $\mathbb{V}_2$ são condicionalmente independentes dado $\mathbb{V}_3$ é relativamente pequeno àquele em que $\mathbb{V}_1$ e $\mathbb{V}_2$ são condicionalmente dependentes. Estudaremos um caso em que é possível observar esta relação em mais detalhe.

Considere que V_1, V_2, e Z são binárias e formam o grafo $V_1 \leftarrow Z \rightarrow V_2$, isto é, Z é um confundidor. Além disso, $\mathbb{P}(Z = 1) = 0.5$, $\mathbb{P}(V_i = 1|Z = j) =: p_j$. Como V_3 é um confundidor, V_1 e V_2 não são d-separados marginalmente. Para quais valores de p temos que V_1 e V_2 são marginalmente independentes? Para que V_1 e V_2 sejam independentes, é necessário que $Cov[V_1, V_2] = 0$. Note que

$$\begin{aligned}
\mathbb{E}[V_i] &= \mathbb{E}[\mathbb{E}[V_i|Z]] = 0.5p_1 + 0.5p_0 \\
\mathbb{E}[V_1V_2] &= \mathbb{E}[|E[V_1V_2|Z]] = 0.5p_1^2 + 0.5p_0^2
\end{aligned}$$

Assim, para que $Cov[V_1, V_2] = 0$, temos:

$$\begin{aligned}
0.5p_1^2 + 0.5p_0^2 &= (0.5p_1 + 0.5p_0)(0.5p_1 + 0.5p_0) \\
0.5p_1^2 + 0.5p_0^2 &= 0.25p_1^2 + +0.5p_1p_0 0.25p_0^2 \\
0.25p_1^2 - 0.5p_1p_0 + 0.25p_0^2 &= 0 \\
0.25(p_1 - p_0)^2 &= 0 \\
p_1 &= p_0
\end{aligned}$$

Em outras palavras, dentre todos (p_0, p_1) no quadrado $[0, 1]^2$, somente os valores no segmento $p_1 = p_0$ tem alguma

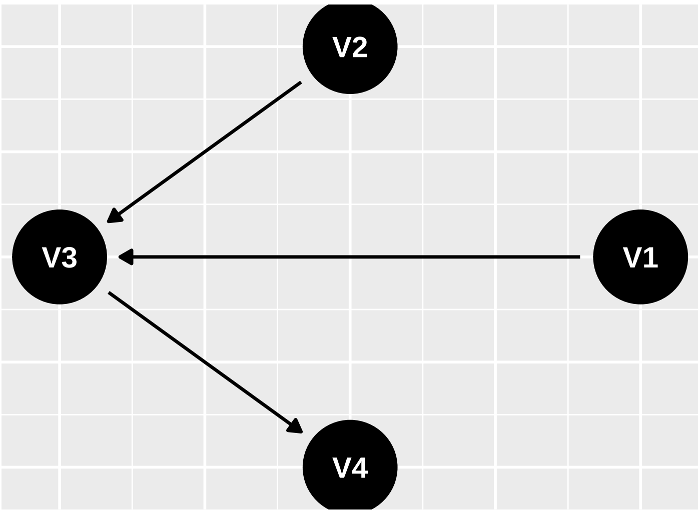

Figura 2.6.: Exemplo em que V4 é um descendente de um colisor, V3.

chance de levarem à independência entre V_1 e V_2. Se imaginarmos que (p_0, p_1) são equidistribuídos em $[0,1]^2$, então a probabilidade de sortearmos valores em que V_1 e V_2 são independentes é 0.

Em conclusão, como V_1 e V_2 não são d-separados, somente para um conjunto pequeno de possíveis f's temos que V_1 e V_2 são independentes.

2.2.3. Exercícios

Exercício 2.43. Considere que f é uma densidade sobre $\mathcal{V} = (V_1, V_2, V_3, V_4)$ que é compatível com o grafo em fig. 2.6. Além disso, cada $V_i \in \{0,1\}$, $V_1, V_2 \sim \text{Bernoulli}(0.5)$, $V_3 \equiv V_1 \cdot V_2$ e $\mathbb{P}(V_4 = i | V_3 = i) = 0.9$, para todo i.

(a) V_1 e V_2 são d-separados dado V_3?

(b) V_1 e V_2 são condicionalmente independentes dado V_3?

(c) V_1 e V_2 são d-separados dado V_4?

(d) V_1 e V_2 são condicionalmente independentes dado V_4?

Exercício 2.44. Prove que se um caminho, $C = (C_1, \ldots, C_n)$, está bloqueado dado $\mathbb{V}$, então sempre que C é um sub-caminho de C^*, isto é, $C^* = (A_1, \ldots, A_m, C_1, \ldots, C_n, B_1, \ldots, B_l)$, temos que C^* está bloqueado dado $\mathbb{V}$.

Exercício 2.45. Prove que se $\mathbb{V}_1 \perp \mathbb{V}_3 | \mathbb{V}_4$ e $\mathbb{V}_2 \perp \mathbb{V}_3 | \mathbb{V}_4$, então $\mathbb{V}_1 \cup \mathbb{V}_2 \perp \mathbb{V}_3 | \mathbb{V}_4$.

Exercício 2.46. Prove que se $\mathbb{V}_1 \perp \mathbb{V}_2 | \mathbb{V}_3$, então para todo $V \in \mathcal{V}$, $V \perp \mathbb{V}_1 | \mathbb{V}_3$ ou $V \perp \mathbb{V}_2 | \mathbb{V}_3$.

Exercício 2.47. Sejam $\mathcal{G}_1 = (\mathcal{V}, \mathcal{E}_1)$ e $\mathcal{G}_2 = (\mathcal{V}, \mathcal{E}_2)$ grafos tais que $\mathcal{E}_1 \subseteq \mathcal{E}_2$. Prove que se $\mathbb{V}_1 \perp^d \mathbb{V}_2 | \mathbb{V}_3$ em $\mathcal{G}_2$, então $\mathbb{V}_1 \perp^d \mathbb{V}_2 | \mathbb{V}_3$ em $\mathcal{G}_1$.

3. Intervenções

3.1. O modelo de probabilidade para intervenções

Com base no modelo estrutural causal discutido no capítulo 2, agora estabeleceremos um significado para o efeito causal de uma variável em outra.

Para iniciar esta discussão, considere as variáveis Z (Sexo), X (Tratamento), e Y (Cura), discutidas no capítulo 1. Podemos considerar que Z é uma causa tanto de X quanto de Y e que X é uma causa de Y. Assim, podemos representar as relações causais entre estas variáveis por meio do grafo na fig. 3.1. Usando este grafo, podemos discutir mais a fundo porque a probabilidade condicional de cura dado tratamento é distinta do efeito causal do tratamento na cura.

Quando calculamos a probabilidade condicional de cura dado o tratamento, estamos perguntando: "Qual é a probabilidade de que um indivíduo selecionado aleatoriamente da população se cure dado que **aprendemos** que recebeu o tratamento?" Para responder a esta pergunta, propagamos a informação do tratamento usado em todos os caminhos do tratamento para a cura. Assim, além do efeito direto que o tratamento tem na cura, o tratamento também está associado ao sexo do paciente, o que indiretamente traz mais informação sobre a cura deste. Isto é, neste caso o tratamento traz informação tanto sobre seus efeitos (cura), quanto sobre suas causas (sexo). Uma outra maneira de verificar estas afirmações é calculando diretamente $f(y|x)$:

$$\begin{aligned} f(y|x) &= \sum_s f(z,y|x) \\ &= \sum_s \frac{f(z,y,x)}{f(x)} \\ &= \sum_s \frac{f(z,x)f(y|z,x)}{f(x)} \\ &= \sum_s f(z|x)f(y|z,x) \end{aligned} \tag{3.1}$$

Notamos na eq. (3.1) que $f(y|x)$ é a média das probabilidades de cura em cada sexo, $f(y|z,x)$, ponderadas pela distribuição do sexo após aprender o tratamento do indivíduo, $f(z|x)$.

A probabilidade condicional de cura dado tratamento não corresponde àquilo que entendemos por efeito causal de tratamento em cura. Este efeito é a resposta para a pergunta: "Qual a probabilidade de que um indivíduo selecionado aleatoriamente da população se cure dado que **prescrevemos** a ele o tratamento?". Ao contrário da primeira pergunta, em que apenas **observamos** a população, nesta segunda fazemos uma **intervenção** sobre o comportamento do indivíduo. Assim, estamos fazendo uma pergunta sobre uma distribuição de probabilidade diferente, em que estamos agindo sobre a unidade amostral. Por exemplo, suponha que prescreveríamos o tratamento a qualquer indivíduo que fosse amostrado. Neste caso, saber qual tratamento foi aplicado não traria qualquer informação sobre o sexo do indivíduo. Em outras palavras, se chamarmos $f(y|do(x))$ como a probabilidade de

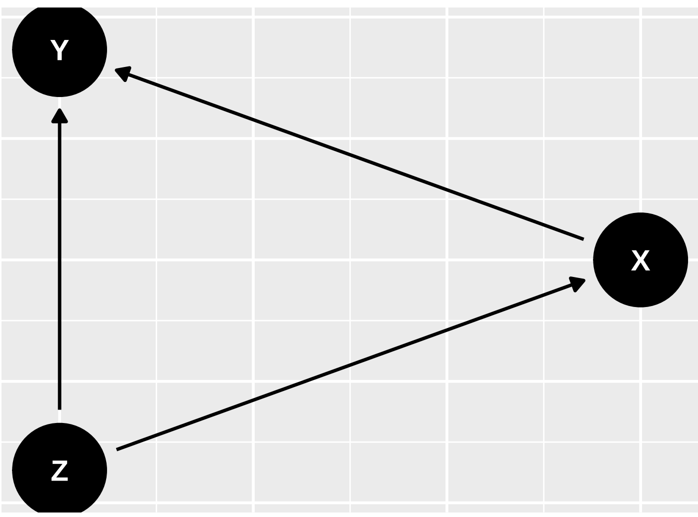

Figura 3.1.: Grafo que representa as relações causais entre Z (Sexo), X (Tratamento), e Y (Cura).

cura dado que fazemos uma intervenção no tratamento, faria sentido obtermos:

$$f(y|do(x)) = \sum_{s} f(z)f(y|z,x) \tag{3.2}$$

Na eq. (3.2) temos que o efeito causal do tratamento na cura é a média ponderada das probabilidades de cura em cada sexo ponderada pelas probabilidades de sexo de um indivíduo retirado aleatoriamente da população. Isto é, ao contrário da eq. (3.1), a distribuição do sexo do indivíduo não é alterada quando fazemos uma intervenção sobre o tratamento.

Com base neste exemplo, podemos generalizar o que entendemos por intervenção. Quando fazemos uma intervenção em uma variável, V_1, tomamos uma ação para que V_1 assuma um determinado valor. Assim, as demais variáveis que comumente seriam causas de V_1 deixam de sê-lo. Por exemplo, para o caso na fig. 3.1, o modelo de intervenção removeria a aresta de Sexo para Tratamento, resultado na fig. 3.2.

Com base nas observações acima, finalmente podemos definir o modelo de probabilidade sob intervenção:

Definição 3.1. Seja $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ um DAG, $(\mathcal{G}, f)$ um SCM (Definição 2.19), e $\mathbb{V}_1 \subseteq \mathcal{V}$. O modelo de probabilidade obtido após uma intervenção em $\mathbb{V}_1$ é dado por:

$$f(\mathcal{V}|do(\mathbb{V}_1)) := \prod_{V_2 \in \mathbb{V}_2} f(V_2|Pa(V_2)) \qquad \text{, ou equivalentemente}$$

$$f(\mathcal{V}|do(\mathbb{V}_1 = \mathbf{v}_1)) := \left(\prod_{(v_1,V_1)\in(\mathbf{v}_1,\mathbb{V}_1)} \mathbb{I}(V_1 = v_1)\right) \cdot \left(\prod_{V_2 \notin \mathbb{V}_1} f(V_2|Pa(V_2))\right)$$

Para compreender a Definição 3.1, podemos comparar o modelo de intervenção com o modelo observacional:

$$f(\mathbb{V}_2|\mathbb{V}_1) \propto f(\mathbb{V}_1, \mathbb{V}_2) = \left(\prod_{V_1 \in \mathbb{V}_1} f(V_1|Pa(V_1)\right) \cdot \left(\prod_{V_2 \in \mathbb{V}_2} f(V_2|Pa(V_2)\right)$$

No modelo observacional, a densidade de $\mathbb{V}_2$ dado $\mathbb{V}_1$ é proporcional ao produto, para todos os vértices, da

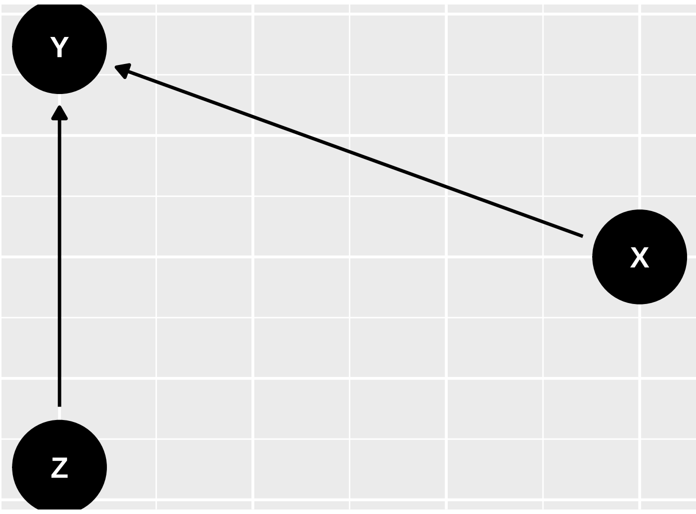

Figura 3.2.: Grafo que representa as relações causais entre S (Sexo), T (Tratamento), e C (Cura) quando há uma intervenção sobre T.

densidade do vértice dadas suas causas. Ao contrário, no modelo de intervenção supomos que os vértices em $\mathbb{V}_1$ são pré-fixados e, assim, não são gerados por suas causas usuais. Assim, na Definição 3.1, a densidade de $\mathbb{V}_2$ dada uma intervenção em $\mathbb{V}_1$ é dada o produto somente nos vértices de $\mathbb{V}_2$ das densidades do vértice dadas suas causas.

Com base na discussão acima, podemos definir o **efeito causal** que um conjunto de variáveis, $\mathbf{X}$, tem em outro conjunto, $\mathbf{Y}$:

Definição 3.2. $\mathbb{E}[\mathbf{Y}|do(\mathbf{X})] := \int \mathbf{y} \cdot f(\mathbf{y}|do(\mathbf{X}))d\mathbf{y}$.

Definição 3.3. O efeito causal médio, ACE,[1] de $X \in \Re$ em $Y \in \Re$ é dado por:

$$ACE = \begin{cases} \mathbb{E}[Y|do(X=1)] - \mathbb{E}[Y|do(X=0)] & \text{, se } X \text{ é binário,} \\ \frac{d\mathbb{E}[Y|do(X=x)]}{dx} & \text{, se } X \text{ é contínuo.} \end{cases}$$

Com a Definição 3.3 podemos finalmente desvendar o Paradoxo de Simpson discutido no capítulo 1. Veremos que o método que desenvolvemos resolve a questão com simplicidade, assim trazendo clareza ao Paradoxo.

Exemplo 3.4. Considere que $(X, Y, Z) \in \Re^3$ são tais que X e Y são as indicadores de que, respectivamente, o paciente recebeu o tratamento e se curou. Além disso, suponha que a distribuição conjunta de (X, Y, Z) é dada

[1] A sigla ACE tem como origem a expressão em inglês, *Average Causal Effect*. Optamos por manter a sigla sem tradução para facilitar a comparação com artigos da área. Em outros contextos, este termo também é chamado de *Average Treatment Effect* e recebe o acrônimo ATE.

pelas frequências na tabela 1.1. Isto é:

$$\begin{aligned}
\mathbb{P}(Z=1) &= \frac{25+55+71+192}{750} \approx 0.46 \\
\mathbb{P}(Z=1|X=0) &= \frac{25+55}{25+55+36+234} \approx 0.23 \\
\mathbb{P}(Z=1|X=1) &= \frac{71+192}{71+192+6+81} \approx 0.75 \\
\mathbb{P}(Y=1|X=0,Z=0) &= \frac{234}{234+36} \approx 0.87 \\
\mathbb{P}(Y=1|X=1,Z=0) &= \frac{81}{81+6} \approx 0.93 \\
\mathbb{P}(Y=1|X=0,Z=1) &= \frac{55}{25+55} \approx 0.69 \\
\mathbb{P}(Y=1|X=1,Z=1) &= \frac{192}{71+192} \approx 0.73
\end{aligned}$$

Agora, veremos que a probabilidade de Y dada uma intervenção em X depende do DAG usado no modelo causal estrutural.

Suponha que Z é a indicadora de que o sexo do paciente é masculino. Neste caso, utilizarem como grafo causal aquele em fig. 3.1. Utilizando este grafo, obtemos:

$$\mathbb{P}_1(Y=i, Z=j|do(X=k)) = \mathbb{P}(Z=j)\mathbb{P}(Y=i|X=k, Z=j) \qquad \text{Definição 3.1} \tag{3.3}$$

Assim,

$$\begin{aligned}
\mathbb{P}_1(Y=1|do(X=1)) &= \mathbb{P}_1(Y=1,Z=0|do(X=1)) + \mathbb{P}_1(Y=1,Z=1|do(X=1)) \\
&= \mathbb{P}(Z=0)\mathbb{P}(Y=1|X=1,Z=0) + \mathbb{P}(Z=1)\mathbb{P}(Y=1|X=1,Z=1) && \text{eq. (3.3)} \\
&\approx 0.54 \cdot 0.93 + 0.46 \cdot 0.73 \approx 0.84 \\
\mathbb{P}_1(Y=1|do(X=0)) &= \mathbb{P}_1(Y=1,Z=0|do(X=0)) + \mathbb{P}_1(Y=1,Z=1|do(X=0)) \\
&= \mathbb{P}(Z=0)\mathbb{P}(Y=1|X=0,Z=0) + \mathbb{P}(Z=1)\mathbb{P}(Y=1|X=0,Z=1) && \text{eq. (3.3)} \\
&\approx 0.54 \cdot 0.87 + 0.46 \cdot 0.69 \approx 0.79
\end{aligned}$$

Portanto, o efeito causal do tratamento na cura quando Z é o sexo do paciente é obtido abaixo:

$$\begin{aligned}
ACE_1 &= \mathbb{E}_1[Y|do(X=1)] - \mathbb{E}_1[Y|do(X=0)] && \text{Definição 3.3} \\
&= \mathbb{P}_1(Y=1|do(X=1)) - \mathbb{P}_1(Y=1|do(X=0)) \approx 0.05 && \text{Definição 3.2}
\end{aligned}$$

Como esperado da discussão na Seção 1.1, o tratamento tem efeito causal médio positivo, isto é, ele aumenta a probabilidade de cura do paciente.

A seguir, consideramos que Z é a indicadora de pressão sanguínea elevada do paciente. Assim, tomamos o grafo causal como aquele na fig. 3.3. Utilizando este grafo, obtemos:

$$\mathbb{P}_2(Y=i, Z=j|do(X=k)) = \mathbb{P}(Z=j|X=k)\mathbb{P}_1(Y=i|X=k, Z=j) \qquad \text{Definição 3.1} \tag{3.4}$$

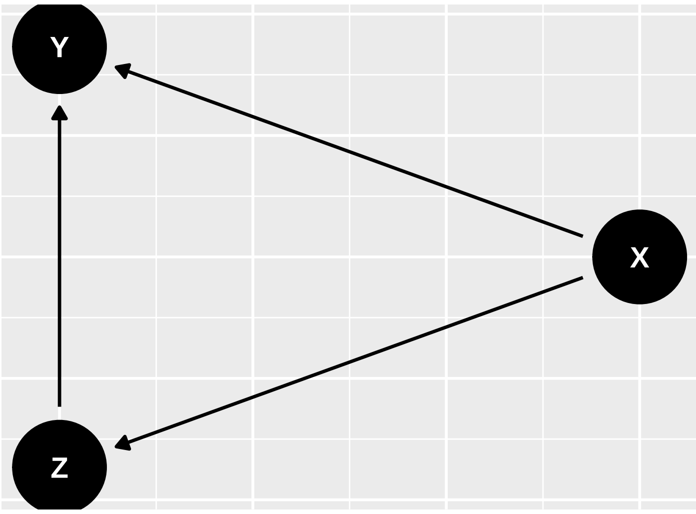

Figura 3.3.: Grafo que representa as relações causais entre Z (Pressão sanguínea elevada), X (Tratamento), e Y (Cura).

Assim,

$$\begin{aligned}
\mathbb{P}_2(Y=1|do(X=1)) &= \mathbb{P}_2(Y=1, Z=0|do(X=1)) + \mathbb{P}_2(Y=1, Z=1|do(X=1)) \\
&= \mathbb{P}(Z=0|X=1)\mathbb{P}(Y=1|X=1, Z=0) + \mathbb{P}(Z=1|X=1)\mathbb{P}(Y=1|X=1, Z=1) \quad \text{eq. (3.4)} \\
&\approx 0.25 \cdot 0.93 + 0.75 \cdot 0.73 \approx 0.78 \\
\mathbb{P}_2(Y=1|do(X=0)) &= \mathbb{P}_2(Y=1, Z=0|do(X=0)) + \mathbb{P}_2(Y=1, Z=1|do(X=0)) \\
&= \mathbb{P}(Z=0|X=0)\mathbb{P}(Y=1|X=0, Z=0) + \mathbb{P}(Z=1|X=0)\mathbb{P}(Y=1|X=0, Z=1) \quad \text{eq. (3.4)} \\
&\approx 0.77 \cdot 0.87 + 0.23 \cdot 0.69 \approx 0.83
\end{aligned}$$

Portanto, o efeito causal do tratamento na cura quando Z é a pressão sanguínea do paciente é obtido abaixo:

$$\begin{aligned}
ACE_1 &= \mathbb{E}_2[Y|do(X=1)] - \mathbb{E}_2[Y|do(X=0)] && \text{Definição 3.3} \\
&= \mathbb{P}_2(Y=1|do(X=1)) - \mathbb{P}_2(Y=1|do(X=0)) \approx -0.05 && \text{Definição 3.2}
\end{aligned}$$

Como esperado da discussão na Seção 1.1, o tratamento tem efeito causal médio negativo, isto é, ele tem como efeito colateral grave a elevação da pressão sanguínea do paciente, reduzindo a probabilidade de cura deste.

Comparando as expressões obtidas em ACE_1 e ACE_2, verificamos que o grafo causal desempenha papel fundamental na determinação do modelo de probabilidade sob intervenção. Ademais, o uso do grafo causal adequado em cada situação formaliza a discussão qualitativa desenvolvida na Seção 1.1. Não há paradoxo!

Além do efeito causal médio, às vezes desejamos determinar o efeito causal de X em Y quando observamos que a unidade amostral faz parte de determinado estrato da população. Em outras palavras, desejamos saber o efeito causal de X em Y quando observamos que outras variáveis, $\mathbf{Z}$, assumem um determinado valor.

Definição 3.5. O efeito causal médio condicional, CACE, de $X \in \Re$ em $Y \in \Re$ dado $\mathbf{Z}$ é:

$$CACE(\mathbf{Z}) = \begin{cases} \mathbb{E}[Y|do(X=1), \mathbf{Z}] - \mathbb{E}[Y|do(X=0), \mathbf{Z}] & \text{, se } X \text{ é binário,} \\ \frac{d\mathbb{E}[Y|do(X=x), \mathbf{Z}]}{dx} & \text{, se } X \text{ é contínuo.} \end{cases}$$

Uma vez estabelecido o modelo de probabilidade utilizado quando estudamos intervenções, agora podemos fazer inferência sobre o efeito causal. Para realizar tal inferência, em geral teremos de abordar duas questões:

1. **Identificação causal**: Temos acesso a dados que são gerados segundo a distribuição observacional. Como é possível determinar o efeito causal em termos da distribuição observacional?

2. **Estimação**: Uma vez estabelecida uma ligação entre a distribuição observacional dos dados e o efeito causal, como é possível estimá-lo?

Nas próximas seções estudaremos algumas estratégias gerais para a resolução destas questões. Consideraremos que desejamos medir o efeito causal de X em Y, onde $X, Y \in \mathcal{V}$.

3.1.1. Exercícios

Exercício 3.6. Considere que X_1 e X_2 são variáveis binárias. Também considere as seguintes definições: **ACE** $:= \mathbb{P}(X_2 = 1|do(X_1 = 1)) - \mathbb{P}(X_2 = 1|do(X_1 = 0))$, e **RD** $:= \mathbb{P}(X_2 = 1|X_1 = 1) - \mathbb{P}(X_2 = 1|X_1 = 0)$. Explique em palavras a diferença entre ACE e RD e apresente um exemplo em que essa diferença ocorre.

Exercício 3.7 (Glymour et al. (2016)[p.32])**.** (X_1, X_2, X_3, X_4) são variáveis binárias tais que X_{i-1} é a única causa imediata de X_i. Além disso, $\mathbb{P}(X_1 = 1) = 0.5$, $\mathbb{P}(X_i = 1|X_{i-1} = 1) = p_{11}$ e $\mathbb{P}(X_i = 1|X_{i-1} = 0) = p_{01}$. Calcule:

(a) $\mathbb{P}(X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 0)$,

(b) $\mathbb{P}(X_4 = 1|X_1 = 1)$, $\mathbb{P}(X_4 = 1|do(X_1 = 1)$,

(c) $\mathbb{P}(X_1 = 1|X_4 = 1)$, $\mathbb{P}(X_1 = 1|do(X_4 = 1)$, e

(d) $\mathbb{P}(X_3 = 1|X_1 = 0, X_4 = 1)$

Exercício 3.8 (Glymour et al. (2016)[p.29])**.** Considere que (U_1, U_2, U_3) são independentes e tais que $U_i \sim N(0, 1)$. Também, $X_1 \equiv U_1$, $X_2 \equiv 3^{-1}X_1 + U_2$, e $X_3 \equiv 2^{-4}X_2 + U_3$. Considere que X_1 é a causa imediata de X_2, que por sua vez é a causa imediata de X_3. Além disso, cada U_i influencia diretamente somente X_i.

(a) Desenhe o DAG que representa a estrutura causal indicada no enunciado.

(b) Calcule $\mathbb{E}[X_2|X_1 = 3]$ e $\mathbb{E}[X_2|do(X_1 = 3)]$.

(c) Calcule $\mathbb{E}[X_3|X_1 = 6]$ e $\mathbb{E}[X_3|do(X_1 = 6)]$.

(d) Calcule $\mathbb{E}[X_1|X_2 = 1]$ e $\mathbb{E}[X_1|do(X_2 = 1)]$.

(e) Calcule $\mathbb{E}[X_2|X_1 = 1, X_3 = 3]$, $\mathbb{E}[X_2|X_1 = 1, do(X_3 = 3)]$, e $\mathbb{E}[X_2|do(X_1 = 1), X_3 = 3]$.

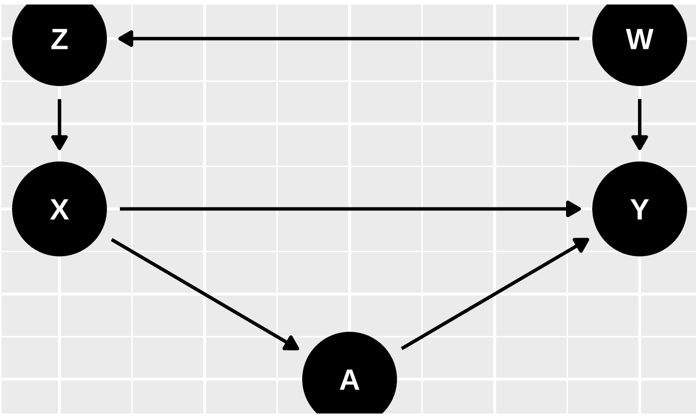

Figura 3.4.: Para medir o efeito causal de X em Y, podemos aplicar o critério backdoor. Neste grafo o único caminho aplicável ao critério backdoor é (X, Z, W, Y). Neste caminho, Z é uma cadeia e W é um confundidor. Assim, todas as possibilidades dentre Z, W, (Z,W) bloqueiam o caminho e satisfazem o critério backdoor.

3.2. Controlando confundidores (critério backdoor)

Um confundidor é uma causa comum, direta ou indireta, de X em Y. Na existência de confundidores, a regressão de Y em X no modelo observacional, $\mathbb{E}[Y|X]$, é diferente desta regressão no modelo de intervenção, $\mathbb{E}[Y|do(X)]$. Isto ocorre pois, quando calculamos $\mathbb{E}[Y|X]$, utilizamos toda a informação em X para prever Y. Esta informação inclui não apenas o efeito causal de X em Y, como também a informação que X traz indiretamente sobre Y pelo fato de ambas estarem associados aos seus confundidores.

Para ilustrar este raciocínio, podemos revisitar o Exemplo 3.4. uma vez que Sexo (Z) é causa comum do Tratamento (X) e da Cura (Y), Z é um confundidor. Quando calculamos $f(y|x)$ (eq. (3.1)), utilizamos não só o efeito direto de X em Y, expresso em $f(y|x,z)$, como também a informação que indireta que X traz sobre Y por meio do confundidor Z, expressa pela combinação de $f(z|x)$ com $f(y|x,z)$.

Esta seção desenvolve uma estratégia para medir o efeito causal chamada de critério *backdoor*, que consiste em bloquear todos os caminhos de informação que passam por confundidores:

Definição 3.9. Seja $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ um grafo causal e $X, Y \in \mathcal{V}$. Dizemos que $\mathbf{Z} \subseteq \mathcal{V} - \{X, Y\}$ satisfaz o critério "backdoor" se:

1. $X \notin Anc(\mathbf{Z})$,

2. Para todo caminho de X em Y, $C = (X, C_2, \ldots, C_{n-1}, Y)$ tal que $(C_2, X) \in \mathcal{E}$, C está bloqueado dado $\mathbf{Z}$.

Exemplo 3.10. No Exemplo 3.4 o único caminho de X em Y em que o vértice ligado a X é pai de X é $X \leftarrow Z \rightarrow Y$. Como Z é um confudidor neste caminho, ele o bloqueia. Assim, Z satisfaz o critério backdoor.

Exemplo 3.11. Considere o grafo causal na fig. 3.4. Para aplicar o critério backdoor, devemos identificar todos os caminhos de X em Y em que o vértice ligado a X é pai de X, isto é, temos $X \leftarrow$. O único caminho deste tipo é: $X \leftarrow Z \leftarrow W \rightarrow Y$. Neste caminho, Z é uma cadeia e W é um confudidor. Assim, é possível bloquear este caminho condicionando em Z, em W, e em (Z, W). Isto é, todos estas combinações satisfazem o critério backdoor.

Exemplo 3.12. Considere o grafo causal na fig. 3.5. Para aplicar o critério backdoor, encontramos todos os caminhos de X em Y em que o vértice ligado a X é pai de X. Há dois caminhos deste tipo: $X \leftarrow A \rightarrow B \rightarrow Y$ e

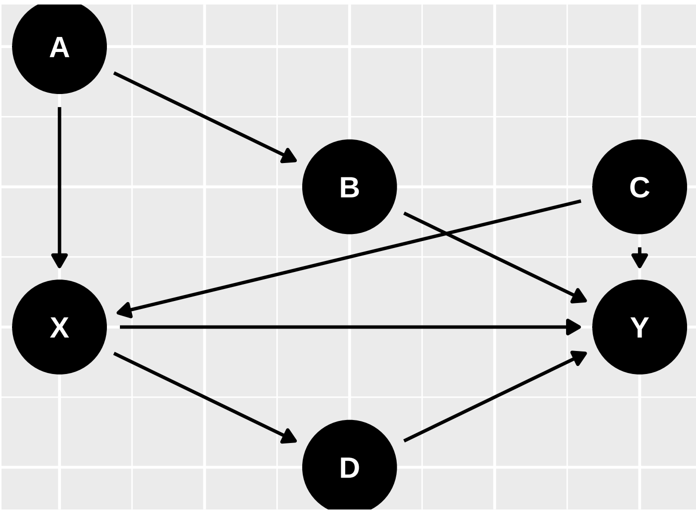

Figura 3.5.: Para medir o efeito causal de X em Y, podemos aplicar o critério backdoor. Neste grafo existem dois caminhos aplicáveis ao critério backdoor: (X, A, B, Y) e (X, C, Y). No primeiro, A é um confundidor. No segundo caminho, C é um confudidor. Assim, (A,C) bloqueia ambos os caminhos e satisfaz o critério backdoor.

$X \leftarrow C \rightarrow Y$. Como A e C são confudidores, respectivamente, no primeiro e segundo caminhos, (A, C) bloqueia ambos eles. Assim (A, C) satisfaz o critério backdoor. Você consegue encontrar outro conjunto de variáveis que satisfaz o critério backdoor?

Também é possível identificar os conjuntos de variáveis que satisfazem o critério backdoor por meio do pacote *dagitty*, como ilustrado a seguir:

```
library(dagitty)
# Especificar o grafo
grafo <- dagitty("dag{
    X[e] Y[o]
    A -> { X B }; B -> { Y }; C -> { X Y };
    X -> { D Y }; D -> Y }")

adjustmentSets(grafo, type = "all")

## { A, C }
## { B, C }
## { A, B, C }
```

O critério backdoor generaliza duas condições especiais que são muito utilizadas. Em uma primeira condição, o valor de X é gerado integralmente por um aleatorizador, independente de todas as demais variáveis. Esta ideia é captada pela Definição 3.13, abaixo:

Definição 3.13. Dizemos que X é um experimento aleatorizado simples se X é ancestral.

Em um experimento aleatorizado simples não há confundidores. Assim, $\emptyset$ satisfaz o critério backdoor:

Lema 3.14. *Se X é um experimento aleatorizado simples, então $\emptyset$ satisfaz o critério backdoor.*

Veremos que em um experimento aleatorizado simples a distribuição intervencional é igual à distribuição observacional. Assim, $\mathbb{E}[Y|do(X)] = \mathbb{E}[Y|X]$ e a inferência causal é reduzida à inferência comumente usadas para a distribuição observacional.

Além disso, o conjunto de todos os pais de X também satisfaz o critério backdoor:

Lema 3.15. $\mathbf{Z} = Pa(X)$ *satisfaz o critério backdoor para medir o efeito causal de X em Y.*

A seguir, veremos como o critério backdoor permite a identificação causal, isto é, uma equivalência entre quantidades de interesse obtidas pelo modelo de intervenção e quantidades obtidas pelo modelo observacional.

3.2.1. Identificação causal usando o critério backdoor

A seguir, o Teorema 3.16 mostra que, se $\mathbf{Z}$ satisfaz o critério backdoor, então é possível ligar algumas distribuições sob intervenção em X a distribuições observacionais:

Teorema 3.16. *Se $\mathbf{Z}$ satisfaz o critério backdoor para medir o efeito causal de X em Y, então*

$$f(\mathbf{z}|do(x)) = f(\mathbf{z}), \ \ e$$
$$f(y|do(x), \mathbf{z}) = f(y|x, \mathbf{z}).$$

O Teorema 3.16 mostra que, se $\mathbf{Z}$ satisfaz o critério backdoor, então distribuição de $\mathbf{Z}$ quando aplicamos uma intervenção em X é igual à distribuição marginal de $\mathbf{Z}$. Além disso, a distribuição condicional de Y dado $\mathbf{Z}$ quando aplicamos uma intervenção em X é igual à distribuição de Y dado X e Z. Assim, o Teorema 3.16 relaciona distribuições que não geraram os dados a distribuições que os geraram. Com base neste resultado, é possível determinar $f(y|do(x))$ a partir de $f(y, x, \mathbf{z})$:

Corolário 3.17. *Se $\mathbf{Z}$ satisfaz o critério backdoor para medir o efeito causal de X em Y, então*

$$f(y|do(x)) = \int f(y|x, \mathbf{z}) f(\mathbf{z}) d\mathbf{z}.$$

Para compreender intuitivamente o Corolário 3.17, podemos retornar ao Exemplo 3.4. Considere o caso em que X, Y, Z são as indicadoras de que, respectivamente, o paciente foi submetido ao tratamento, se curou e, é de sexo masculino. Similarmente ao Teorema 3.16, vimos em Exemplo 3.4 que $f(y|do(x))$ é a média de $f(y|x, z)$ ponderada por $f(z)$. Nesta ponderação, utilizamos $f(z)$ ao invés de $f(z|x)$ pois Z é um confundidor e, assim, no modelo intervencional não propagamos a informação em X por esta variável. A mesma lógica se aplica às variáveis que satisfazem o critério backdoor.

Para calcular quantidades como o ACE (Definição 3.3), utilizamos $\mathbb{E}[Y|do(X)]$. Por meio do Teorema 3.16, é possível obter equivalências entre $\mathbb{E}[Y|do(X)]$ e esperanças obtidas no modelo observacional. Estas equivalências são descritas nos Teoremas 3.18 e 3.19.

Teorema 3.18. *Se $\mathbf{Z}$ satisfaz o critério backdoor para medir o efeito causal de X em Y, então*

$$\mathbb{E}[g(Y)|do(X = x), \mathbf{Z}] = \mathbb{E}[g(Y)|X = x, \mathbf{Z}], \ \ e$$
$$\mathbb{E}[g(Y)|do(X = x)] = \mathbb{E}[\mathbb{E}[g(Y)|X = x, \mathbf{Z}]]$$

Teorema 3.19. *Se $\mathbf{Z}$ satisfaz o critério backdoor para medir o efeito causal de X em Y e X é discreto, então*

$$\mathbb{E}[g(Y)|do(X=x),\mathbf{Z}] = \frac{\mathbb{E}[g(Y)\mathbb{I}(X=x)|\mathbf{Z}]}{f(x|\mathbf{Z})}, \text{ e}$$
$$\mathbb{E}[g(Y)|do(X=x)] = \mathbb{E}\left[\frac{g(Y)\mathbb{I}(X=x)}{f(x|\mathbf{Z})}\right]$$

A seguir, veremos como os Teoremas 3.18 e 3.19 podem ser usados para estimar o efeito causal. Para provar resultados sobre os estimadores obtidos, a seguinte definição será útil

Definição 3.20. Seja $\hat{g}$ um estimador treinado com os dados $(\mathcal{V}_1, \ldots, \mathcal{V}_n)$. Dizemos que $\hat{g}$ é invariante a permutações se o estimador não depende da ordem dos dados. Isto é, para qualquer permutações dos índices, $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$, $\hat{g}(\mathcal{V}_1, \ldots, \mathcal{V}_n) \equiv \hat{g}(\mathcal{V}_{\pi(1)}, \ldots, \mathcal{V}_{\pi(n)})$

Exemplo 3.21. A média amostral é invariante a permutações pois, para qualquer permutação π,

$$\frac{\sum_{i=1}^n X_i}{n} = \frac{\sum_{i=1}^n X_{\pi(i)}}{n}.$$

3.2.2. Estimação usando o critério backdoor

Fórmula do ajuste

O Teorema 3.18 determina que, se $\mathbf{Z}$ satisfaz o critério backdoor, então $\mathbb{E}[Y|do(X),\mathbf{Z}] = \mathbb{E}[Y|X,\mathbf{Z}]$. Como $\mu(X,Z) := \mathbb{E}[Y|X,\mathbf{Z}]$ é a função de regressão de Y em X e Z, podemos estimar μ utilizando quaisquer métodos de estimação para regressão. Por exemplo, se Y é contínua, possíveis métodos são: regressão linear, Nadaraya-Watson, floresta aleatória de regressão, redes neurais, ... Por outro lado, se Y é discreta, então a função de regressão é estimada por métodos de classificação como: regressão logística, k-NN, floresta aleatória de classificação, redes neurais, ... Para qualquer opção escolhida, denotamos o estimador de μ por $\widehat{\mu}$.

Utilizando $\widehat{\mu}$, podemos estimar $CACE(\mathbf{Z})$ diretamente. Para tal, note que $CACE(\mathbf{Z})$ é função de $\mathbb{E}[Y|do(X),\mathbf{Z}]$. Como o Teorema 3.18 garante que $\mathbb{E}[Y|do(X=x),\mathbf{Z}] = \mu(x,\mathbf{Z})$, podemos definir o estimador

$$\widehat{\mathbb{E}}_1[Y|do(X=x),\mathbf{Z}] := \widehat{\mu}(x,\mathbf{Z}).$$

O Teorema 3.18 também orienta a estimação do ACE. Similarmente ao caso anterior, o ACE é função de $\mathbb{E}[Y|do(X)]$. Pelo Teorema 3.18, $\mathbb{E}[Y|do(X=x)] = \mathbb{E}[\mu(x,\mathbf{Z})]$. Assim, se $\widehat{\mu} \approx \mu$, $\mathbb{E}[Y|do(X=x)] \approx \mathbb{E}[\widehat{\mu}(x,\mathbf{Z})]$. Como $\mathbb{E}[\widehat{\mu}(x,\mathbf{Z})]$ é simplesmente uma média populacional, podemos estimá-la com base na média amostral:

$$\widehat{\mathbb{E}}_1[Y|do(X=x)] := \frac{\sum_{i=1}^n \widehat{\mu}(x,\mathbf{Z}_i)}{n} \approx \mathbb{E}[\widehat{\mu}(x,\mathbf{Z})]$$

Definição 3.22. Considere que $\mathbf{Z}$ satisfaz o critério backdoor para medir o efeito causal de X em Y e $\widehat{\mu}(x,\mathbf{z})$ é uma estimativa da regressão $\mathbb{E}[Y|X=x,\mathbf{Z}=\mathbf{z}]$. Os estimadores de $\mathbb{E}[Y|do(X=x),\mathbf{Z}]$ e $\mathbb{E}[Y|do(X=x)]$ pela fórmula do ajuste são:

$$\widehat{\mathbb{E}}_1[Y|do(X=x),\mathbf{Z}] := \widehat{\mu}(x,\mathbf{Z})$$
$$\widehat{\mathbb{E}}_1[Y|do(X=x)] := \frac{\sum_{i=1}^n \widehat{\mu}(x,\mathbf{Z}_i)}{n}$$

A seguir mostraremos que, se $\widehat{\mu}$ converge para μ, então $\widehat{\mathbb{E}}_1[Y|do(X=x)]$ converge para $\mathbb{E}[Y|do(X=x)]$. Em

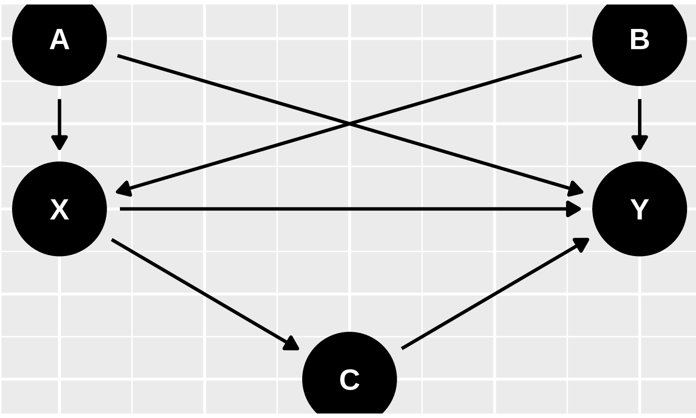

Figura 3.6.: DAG usado como exemplo para estimar efeito de X em Y.

outras palavras, é possível utilizar $\widehat{\mathbb{E}}_1[Y|do(X = x)]$ para estimar o efeito causal de X em Y por meio de expressões como o ACE.

Teorema 3.23. *Seja* $\mu(X, \mathbf{Z}) := \mathbb{E}[Y|X, \mathbf{Z}]$. *Se* $\mathbf{Z}$ *satisfaz o critério backdoor para medir o efeito causal de* X *em* Y, $\mathbb{E}[|\mu(x, \mathbf{Z}_1)|] < \infty$, $\mathbb{E}[|\widehat{\mu}(x, \mathbf{Z}_1) - \mu(x, \mathbf{Z}_1)|] = o(1)$, *e* $\widehat{\mu}$ *é invariante a permutações (Definição 3.20), então* $\widehat{\mathbb{E}}_1[Y|do(X = x)] \xrightarrow{\mathbb{P}} \mathbb{E}[Y|do(X = x)]$.

A seguir, utilizamos dados simulados para ilustrar a implementação da fórmula do ajuste.

Exemplo 3.24. Considere que o grafo causal é dado pela fig. 3.6. Vamos supor que os dados são gerados da seguinte forma: $\sigma^2 = 0.01$, $A \sim N(0, \sigma^2)$, $B \sim N(0, \sigma^2)$, $\epsilon \sim Bernoulli(0.95)$ $X \equiv \mathbb{I}(A + B > 0)\epsilon + \mathbb{I}(A + B < 0)(1 - \epsilon)$, $C \sim N(X, \sigma^2)$, e $Y \sim N(A + B + C + X, \sigma^2)$:

```r
# Especificar o grafo
grafo <- dagitty::dagitty("dag {
    X[e] Y[o]
    {A B} -> { X Y }; X -> {C Y}; C -> Y }")

# Simular os dados
n <- 10^5
sd = 0.1
A <- rnorm(n, 0, sd)
B <- rnorm(n, 0, sd)
eps <- rbinom(n, 1, 0.8)
X <- as.numeric(eps*((A + B) > 0) +
                  (1-eps)*((A + B) <= 0))
C <- rnorm(n, X, sd)
Y <- rnorm(n, A + B + C + X, sd)
data <- dplyr::tibble(A, B, C, X, Y)
```

Estimaremos o efeito causal pela fórmula do ajuste (Definição 3.22). Iniciaremos a análise utilizando $\widehat{\mu}$ como sendo uma regressão linear simples:

```r
# Sejam Z variáveis que satisfazem o critério backdoor para
# estimar o efeito causal de causa em efeito em grafo.
# Retorna uma fórmula do tipo Y ~ X + Z_1 + ... + Z_d
fm_ajuste <- function(grafo, causa, efeito)
{
  var_backdoor <- dagitty::adjustmentSets(grafo)[[1]]
  regressores = c(causa, var_backdoor)
  fm = paste(regressores, collapse = "+")
  fm = paste(c(efeito, fm), collapse = "~")
  as.formula(fm)
}

# Estima E[Efeito|do(causa = x)] pela
# formula do ajuste usando mu_chapeu como regressao
est_do_x_lm <- function(data, mu_chapeu, causa, x)
{
  data %>%
    dplyr::mutate({{causa}} := x) %>%
    predict(mu_chapeu, newdata = .) %>%
    mean()
}

# Estimação do ACE com regressão linear simples
fm <- fm_ajuste(grafo, "X", "Y")
mu_chapeu_lm <- lm(fm, data = data)
ace_ajuste_lm = est_do_x_lm(data, mu_chapeu_lm, "X", 1) -
  est_do_x_lm(data, mu_chapeu_lm, "X", 0)
round(ace_ajuste_lm)

## [1] 2
```

Em alguns casos, não é razoável supor que $\mathbb{E}[Y|X, \mathbf{Z}]$ é linear. Nestas situações, é fácil adaptar o código anterior para algum método não-paramétrico arbitrário. Exibimos uma implementação usando XGBoost (Chen et al., 2023):

```r
library(xgboost)
var_backdoor <- dagitty::adjustmentSets(grafo, "X", "Y")[[1]]
mu_chapeu <- xgboost(
  data = data %>%
    dplyr::select(all_of(c(var_backdoor, "X"))) %>%
    as.matrix(),
  label = data %>%
    dplyr::select(Y) %>%
    as.matrix(),
```

```r
  nrounds = 100,
  objective = "reg:squarederror",
  early_stopping_rounds = 3,
  max_depth = 2,
  eta = .25,
  verbose = FALSE
)

est_do_x_xgb <- function(data, mu_chapeu, causa, x)
{
  data %>%
    dplyr::mutate({{causa}} := x) %>%
    dplyr::select(c(var_backdoor, causa)) %>%
    as.matrix() %>%
    predict(mu_chapeu, newdata = .) %>%
    mean()
}

ace_est_xgb = est_do_x_xgb(data, mu_chapeu, "X", 1) -
  est_do_x_xgb(data, mu_chapeu, "X", 0)
round(ace_est_xgb, 2)

## [1] 2
```

Como o modelo linear era adequado para $\mathbb{E}[Y|X, \mathbf{Z}]$, não vemos diferença entre a estimativa obtida pela regressão linear simples e pelo XGBoost. Mas será que as estimativas estão adequadas? Como simulamos os dados, é possível calcular diretamente $\mathbb{E}[Y|do(X = x)]$:

$$
\begin{aligned}
\mathbb{E}[Y|do(X = x)] &= \mathbb{E}[\mathbb{E}[Y|X = x, A, B]] && \text{Teorema 3.18} \\
&= \mathbb{E}[\mathbb{E}[\mathbb{E}[Y|X = x, A, B, C]|X = x, A, B]] && \text{Lei da esperança total} \\
&= \mathbb{E}[\mathbb{E}[A + B + C + X|X = x, A, B]] && Y \sim N(A + B + C + X, \sigma^2) \\
&= \mathbb{E}[A + B + 2x] && C \sim N(X, \sigma^2) \\
&= 2x && \mathbb{E}[A] = \mathbb{E}[B] = 0
\end{aligned} \tag{3.5}
$$

Uma vez calculado $\mathbb{E}[Y|do(X = x)]$, podemos obter o ACE:

$$
\begin{aligned}
ACE &= \mathbb{E}[Y|do(X = 1)] - \mathbb{E}[Y|do(X = 0)] \\
&= 2 \cdot 1 - 2 \cdot 0 = 2 && \text{eq. (3.5)}
\end{aligned}
$$

Portanto, as estimativas do ACE obtidas pela regressão linear e pelo xgboost estão adequadas.

Ponderação pelo inverso do escore de propensidade (IPW)

Uma outra forma de estimar $\mathbb{E}[Y|do(X=x)]$ e $\mathbb{E}[Y|do(X=x),\mathbf{Z}]$ é motivada pelo Teorema 3.19. Este resultado determina que, se $\mathbf{Z}$ satisfaz o critério backdoor, então

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

Na segunda expressão, $\mathbb{E}[Y\mathbb{I}(X=x)|\mathbf{Z}]$ é a regressão de $Y\mathbb{I}(X=x)$ em $\mathbf{Z}$. Assim, esta quantidade pode ser estimada por um método de regressão arbitrário, que denotaremos por $\widehat{\mathbb{E}}[Y\mathbb{I}(X=x)|\mathbf{Z}]$. Também $f(x|\mathbf{z})$ é usualmente chamado de *escore de propensidade*. Este escore captura a forma como os confundidores atuam sobre X nos dados observacionais. Como f em geral é desconhecido, $f(x|\mathbf{z})$ também o é. Contudo, quando X é discreto, podemos estimar $f(x|\mathbf{z})$ utilizando algum algoritmo arbitrário de classificação. Denotaremos esta estimativa por $\widehat{f}(x|\mathbf{z})$. Se a estimativa for boa, temos

$$\mathbb{E}[Y|do(X=x),\mathbf{Z}] = \frac{\widehat{\mathbb{E}}[Y\mathbb{I}(X=x)|\mathbf{Z}]}{f(x|\mathbf{Z})} \approx \frac{\widehat{\mathbb{E}}[Y\mathbb{I}(X=x)|\mathbf{Z}]}{\widehat{f}(x|\mathbf{z})}.$$

O Teorema 3.19 também orienta a estimação de $\mathbb{E}[Y|do(X=x)]$. Se $\mathbf{Z}$ satisfaz o critério backdoor, então

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

Como nesta expressão a esperança é uma média populacional, ela pode ser aproximada pela média amostral

$$\mathbb{E}\left[\frac{Y\mathbb{I}(X=x)}{f(x|\mathbf{Z})}\right] \approx n^{-1}\sum_{i=1}^{n}\frac{Y_i\mathbb{I}(X_i=x)}{\widehat{f}(x|\mathbf{Z}_i)}.$$

Combinando estas aproximações, obtemos:

Definição 3.25. Considere que $\mathbf{Z}$ satisfaz o critério backdoor para medir o efeito causal de X em Y e $\widehat{f}(x|\mathbf{z})$ é uma estimativa de $f(x|\mathbf{z})$. Os estimadores de $\mathbb{E}[Y|do(X=x),\mathbf{Z}]$ e $\mathbb{E}[Y|do(X=x)]$ por IPW são:

$$\widehat{\mathbb{E}}_2[Y|do(X=x),\mathbf{Z}] := \frac{\widehat{\mathbb{E}}[Y\mathbb{I}(X=x)|\mathbf{Z}]}{\widehat{f}(x|\mathbf{z})}$$
$$\widehat{\mathbb{E}}_2[Y|do(X=x)] := n^{-1}\sum_{i=1}^{n}\frac{Y_i\mathbb{I}(X_i=x)}{\widehat{f}(x|\mathbf{Z}_i)}.$$

Se $\widehat{f}$ converge para f, então sob condições relativamente pouco restritivas $\widehat{\mathbb{E}}_2[Y|do(X=x)]$ converge para $\mathbb{E}[Y|do(X=x)]$.

Teorema 3.26. *Se $\widehat{f}$ é invariante a permutações (Definição 3.20), $\mathbb{E}[|\widehat{f}(x|\mathbf{Z}_1)-f(x|\mathbf{Z}_1)|]=o(1)$, e existe $M>0$ tal que $\sup_{\mathbf{z}}\mathbb{E}[|Y|\mathbb{I}(X=x)|\mathbf{Z}=\mathbf{z}]<M$, e existe $\delta>0$ tal que $\inf_z \min\{f(x|\mathbf{Z}_1),\widehat{f}(x|\mathbf{Z}_1)\}>\delta$, então $\widehat{\mathbb{E}}_2[Y|do(X=x)] \xrightarrow{\mathbb{P}} \mathbb{E}[Y|do(X=x)]$.*

A seguir, utilizamos novamente dados simulados para ilustrar a implementação de IPW:

Exemplo 3.27. Considere que o grafo causal e o modelo de geração dos dados são idênticos àqueles do Exemplo 3.24. Iniciaremos a análise utilizando regressão logística para estimar $\widehat{f}$.

```r
# Sejam Z variáveis que satisfazem o critério backdoor para
# estimar o efeito causal de causa em efeito em grafo.
# Retorna uma fórmula do tipo X ~ Z_1 + ... + Z_d
fm_ipw <- function(grafo, causa, efeito)
{
 var_backdoor <- dagitty::adjustmentSets(grafo)[[1]]
 fm = paste(var_backdoor, collapse = "+")
 fm = paste(c(causa, fm), collapse = "~")
 as.formula(fm)
}

# Estimação do ACE por IPW onde
# Supomos X binário e
# f_1 é o vetor P(X_i=1|Z_i)
ACE_ipw <- function(data, causa, efeito, f_1)
{
 data %>%
 mutate(f_1 = f_1,
        est_1 = {{efeito}}*({{causa}}==1)/f_1,
        est_0 = {{efeito}}*({{causa}}==0)/(1-f_1)
 ) %>%
 summarise(do_1 = mean(est_1),
           do_0 = mean(est_0)) %>%
 mutate(ACE = do_1 - do_0) %>%
 dplyr::select(ACE)
}

fm <- fm_ipw(grafo, "X", "Y")
f_chapeu <- glm(fm, family = "binomial", data = data)
f_1_lm <- predict(f_chapeu, type = "response")
ace_ipw_lm <- data %>% ACE_ipw(X, Y, f_1_lm) %>% as.numeric()
ace_ipw_lm %>% round(2)

## [1] 2.09
```

Também é fácil adaptar o código acima para estimar ACE por IPW utilizando algum método não-paramétrico para estimar $\widehat{f}$. Abaixo há um exemplo utilizando o XGBoost:

```r
var_backdoor <- dagitty::adjustmentSets(grafo)[[1]]
f_chapeu <- xgboost(
 data = data %>%
   dplyr::select(all_of(var_backdoor)) %>%
   as.matrix(),
 label = data %>%
```

```
  dplyr::select(X) %>%
  as.matrix(),
 nrounds = 100,
 objective = "binary:logistic",
 early_stopping_rounds = 3,
 max_depth = 2,
 eta = .25,
 verbose = FALSE
)

covs <- data %>% dplyr::select(all_of(var_backdoor)) %>% as.matrix()
f_1 <- predict(f_chapeu, newdata = covs)
data %>% ACE_ipw(X, Y, f_1) %>% as.numeric() %>% round(2)

## [1] 1.97
```

Estimador duplamente robusto

Os Teoremas 3.23 e 3.26 mostram que, sob suposições diferentes, $\widehat{\mathbb{E}}_1[Y|do(X=x)]$ e $\widehat{\mathbb{E}}_2[Y|do(X=x)]$ convergem para $\mathbb{E}[Y|do(X=x)]$. A ideia do estimador duplamente robusto é combinar ambos os estimadores de forma a garantir esta convergência sob suposições mais fracas. Para tal, a ideia por trás do estimador duplamente é que este convirja junto a $\widehat{\mathbb{E}}_1[Y|do(X=x)]$ quando este é consistente e para $\widehat{\mathbb{E}}_2[Y|do(X=x)]$ quando aquele o é.

Definição 3.28. Sejam $\mathbf{Z}$ variáveis que satisfazem o critério backdoor para medir o efeito causal de X em Y e sejam $\widehat{f}$ e $\widehat{\mu}$ tais quais nas Definições 3.22 e 3.25. O estimador duplamente robusto para $\mathbb{E}[Y|do(X=x)]$, $\widehat{\mathbb{E}}_3[Y|do(X=x)]$ é tal que

$$\widehat{\mathbb{E}}_3[Y|do(X=x)] = \widehat{\mathbb{E}}_1[Y|do(X=x)] + \widehat{\mathbb{E}}_2[Y|do(X=x)] - \sum_{i=1}^{n} \frac{\mathbb{I}(X_i=x)\widehat{\mu}(x,\mathbf{Z}_i)}{n\hat{f}(x|\mathbf{Z}_i)}$$

O estimador duplamente robusto é consistente para $\mathbb{E}[Y|do(X=x)]$ tanto sob as condições do Teorema 3.23 quanto sob as do Teorema 3.26. A ideia básica é que, sob as condições do Teorema 3.23, $\widehat{\mathbb{E}}_1[Y|do(X=x)]$ é consistente para $\mathbb{E}[Y|do(X=x)]$ e $\widehat{\mathbb{E}}_2[Y|do(X=x)] - \sum_{i=1}^{n} \frac{\mathbb{I}(X_i=x)\widehat{\mu}(x,\mathbf{Z}_i)}{n\hat{f}(x|\mathbf{Z}_i)}$ converge para 0. Isto é, quando $\widehat{\mathbb{E}}_1[Y|do(X=x)]$ é consistente, o estimador duplamente robusto seleciona este termo. Similarmente, sob as condições do Teorema 3.26, $\widehat{\mathbb{E}}_2[Y|do(X=x)]$ é consistente para $\mathbb{E}[Y|do(X=x)]$ e $\widehat{\mathbb{E}}_1[Y|do(X=x)] - \sum_{i=1}^{n} \frac{\mathbb{I}(X_i=x)\widehat{\mu}(x,\mathbf{Z}_i)}{n\hat{f}(x|\mathbf{Z}_i)}$ converge para 0.

Teorema 3.29. *Suponha que existe $\epsilon > 0$ tal que $\inf_{\mathbf{z}} \widehat{f}(x|\mathbf{z}) > \epsilon$, existe $M > 0$ tal que $\sup_{\mathbf{z}} \widehat{\mu}(x,\mathbf{z}) < M$, e $\widehat{\mu}$ e $\widehat{f}$ são invariantes a permutações (Definição 3.20). Se as condições do Teorema 3.23 ou do Teorema 3.26 estão satisfeitas, então*

$$\widehat{\mathbb{E}}_3[Y|do(X=x)] \xrightarrow{\mathbb{P}} \mathbb{E}[Y|do(X=x)].$$

Exemplo 3.30 (Estimador duplamente robusto)**.** Considere que o grafo causal e o modelo de geração dos dados são iguais àqueles descritos no Exemplo 3.24. Para implementar o estimador duplamente robusto combinaremos o estimador da fórmula do ajuste obtido por regressão linear no Exemplo 3.24 e aquele de IPW por regressão logística no Exemplo 3.27.

```
mu_1_lm <- data %>%
  dplyr::mutate(X = 1) %>%
  predict(mu_chapeu_lm, newdata = .)
mu_0_lm <- data %>%
  dplyr::mutate(X = 0) %>%
  predict(mu_chapeu_lm, newdata = .)
corr <- data %>%
  mutate(mu_1 = mu_1_lm,
         mu_0 = mu_0_lm,
         f_1 = f_1_lm,
         corr_1 = (X == 1)*mu_1/f_1,
         corr_0 = (X == 0)*mu_0/(1-f_1)) %>%
  summarise(corr_1 = mean(corr_1),
            corr_0 = mean(corr_0)) %>%
  mutate(corr = corr_1 - corr_0) %>%
  dplyr::select(corr) %>%
  as.numeric()
ace_rob_lm <- ace_ajuste_lm + ace_ipw_lm - corr
ace_rob_lm %>% round(2)

## [1] 2
```

3.2.3. Exercícios

Exercício 3.31. Prove o Lema 3.14.

Exercício 3.32. Prove o Lema 3.15.

Exercício 3.33. Prove que a variância amostral satisfaz o Definição 3.20.

Exercício 3.34. Utilizando como referência o grafo e o código no Exemplo 3.24, simule dados tais que a estimativa do ACE é diferente quando um método de regressão linear e um de regressão não-paramétrica são usados.

3.2.4. Regression Discontinuity Design (RDD)

Em determinadas situações, X é completamente determinado pelos confundidores, $\mathbf{Z}$ (Lee and Lemieux, 2010). Por exemplo, considere que desejamos determinar o efeito causal que um determinado programa social do governo traz sobre o nível de educação dos cidadãos. Neste caso, X é a indicadora de que o indíviduo é elegível ao programa e Y mede o seu nível de educação. Em alguns casos, é razoável supor que X é completamente determinado por Z, a renda do indivíduo.

A situação acima traz desafios para a fórmula do ajuste e IPW discutidos anteriormente. Primeiramente, como $X = h(\mathbf{Z})$, não é possível estimar $\mathbb{E}[Y|X = x, \mathbf{Z} = \mathbf{z}]$ quando $x \neq h(\mathbf{z})$. Portanto, não é possível utilizar a fórmula do ajuste, uma vez que ela se baseia na expressão $\mathbb{E}[\mathbb{E}[Y|X = x, \mathbf{Z}]]$. Similarmente, o estimador de IPW envolve uma divisão por $f(x|\mathbf{Z})$. Assim, quando $x \neq h(\mathbf{Z})$ há uma divisão por 0, o que torna o estimador indefinido.

Identificação causal no RDD

Apesar destas dificuldades, é possível medir nestas situações parte do efeito causal de X em Y. Suponha que $\mathbf{Z} \in \Re$ e que existe z_1 tal que $X = \mathbb{I}(\mathbf{Z} \geq \mathbf{z}_1)$. Por exemplo, um benefício pode estar disponível apenas para cidadãos que tenham renda abaixo de um teto ou uma lei pode ter efeitos a partir de uma determinada data.

Neste caso, podemos estar interessados em $\mathbb{E}[Y|do(X = x), \mathbf{Z} = \mathbf{z}_1]$, o efeito causal que X tem na fronteira de sua implementação. Intuitivamente, próximo a esta fronteira, as unidades amostrais são todas similares em relação aos confundidores. Assim, se na fronteira houver uma diferença em Y entre os valores de X, esta diferença deve ser decorrente do efeito causal de X. Esta intuição é formalizada no resultado de identicação causal abaixo:

Teorema 3.35 (Hahn et al. (2001))**.** *Considere que $\mathbf{Z} \in \Re$ satisfaz o critério backdoor para estimar o efeito causal de $X \in \{0, 1\}$ em Y e que $\mathbb{E}[Y|do(X = 0), \mathbf{Z}]$ e $\mathbb{E}[Y|do(X = 1), \mathbf{Z}]$ são contínuas em $\mathbf{Z} = \mathbf{z}_1$.*

Se $X \equiv \mathbb{I}(\mathbf{Z} \geq \mathbf{z}_1)$, então

$$CACE(\mathbf{Z} = \mathbf{z}_1) = \lim_{\mathbf{z} \downarrow \mathbf{z}_1} \mathbb{E}[Y|\mathbf{Z} = \mathbf{z}] - \lim_{\mathbf{z} \uparrow \mathbf{z}_1} \mathbb{E}[Y|\mathbf{Z} = \mathbf{z}].$$

Se $f(x|\mathbf{Z}) \in (0, 1)$ é contínua exceto em $\mathbf{z}_1$, então

$$CACE(\mathbf{Z} = \mathbf{z}_1) = \frac{\lim_{\mathbf{z} \downarrow \mathbf{z}_1} \mathbb{E}[Y|\mathbf{Z} = \mathbf{z}] - \lim_{\mathbf{z} \uparrow \mathbf{z}_1} \mathbb{E}[Y|\mathbf{Z} = \mathbf{z}]}{\lim_{\mathbf{z} \downarrow \mathbf{z}_1} f(X = 1|\mathbf{Z} = \mathbf{z}) - \lim_{\mathbf{z} \uparrow \mathbf{z}_1} f(X = 1|\mathbf{Z} = \mathbf{z})}.$$

Um detalhe sutil do Teorema 3.35 é que $X \equiv \mathbb{I}(\mathbf{Z} > \mathbf{z}_1)$ não é o suficiente para termos certeza que $\mathbf{Z}$ satisfaz o critério backdoor. Por exemplo, considere que o governo criasse um benefício fiscal para todas empresas sediadas em um determinado município. Neste caso, a ocorrência do benefício é função da sede da empresa. Contudo, a relação causal é mais complexa. Se o benefício for suficientemente alto, poderia motivar empresas a moverem sua sede para o município. Em outras palavras, o benefício seria causa da localização da sede e não o contrário. Neste caso, não seria possível aplicar o Teorema 3.35. Este tipo de raciocínio indica que a análise por RDD é mais efetiva quando é difícil interferir sobre o valor de $\mathbf{Z}$. Por exemplo, como um indivíduo não pode interferir sobre a sua idade, é mais fácil justificar o uso de RDD em uma campanha de vacinação em que apenas indivíduos acima de uma determinada idade são vacinados.

Um outro ponto importante de interpretação do Teorema 3.35 é que, embora $\mathbb{E}[Y|do(X = 0), \mathbf{Z}]$ e $\mathbb{E}[Y|do(X = 1), \mathbf{Z}]$ sejam supostas contínuas, $f(X = 1|\mathbf{Z})$ e $\mathbb{E}[Y|\mathbf{Z}]$ não o são. Intuitivamente, podemos imaginar X representa a indicadora de que uma determinada política é adotada. Por exemplo, podemos imaginar que X indica que um indivíduo foi vacinado, $\mathbf{Z}$ a sua idade e Y a sua hospitalização. Neste caso, $\mathbb{E}[Y|do(X = 0), \mathbf{Z}]$ e $\mathbb{E}[Y|do(X = 1), \mathbf{Z}]$ representam a taxa de hospitalização quando todos os indivíduos são vacinados ou quando todos eles não o são. Nestas situações, seria razoável supor que a taxa de hospitalização é contínua em função da idade, pois não esperamos que exista uma grande descontinuidade nas condições de saúde entre indivíduos com 69 e com 70 anos de idade. Este tipo de conclusão muitas vezes é resumido pela expressão em latim *natura non facit saltus* (a natureza não faz saltos). Por outro lado, nos dados observados, a política não é adotada para uma faixa de valores de $\mathbf{Z}$ e passa a ser adotada a partir de um ponto, o que é responsável pela descontinuidade em $\mathbb{E}[Y|\mathbf{Z}]$ e em $f(X = 1|\mathbf{Z})$. Podemos imaginar que a vacinação é empregada somente em indivíduos com mais de 70 anos. Esta descontinuidade na política humana cria uma diferença importante entre indivíduos com 69 e com 70 anos, o que explica uma diferença grande nas taxas de hospitalização entre estas idades nos dados observados.

Estimação no RDD

O Teorema 3.35 indica que $CACE(\mathbf{Z} = \mathbf{z}_1)$ é função da regressão de Y sobre $\mathbf{Z}$, $\mathbb{E}[Y|\mathbf{Z}]$, e sobre o classificador, $f(X = 1|\mathbf{Z})$. Uma possível estratégia é estimarmos estas quantidades separadamente e, a seguir, estimarmos o $CACE$ trocando as quantias populacionais pelas quantias estimadas.

Uma dificuldade nesta estratégia é que sabemos que $\mathbb{E}[Y|\mathbf{Z}]$ e $f(X = 1|\mathbf{Z})$ são discontínuas. Para lidar com esta dificuldade, uma possibilidade é realizar uma regressão para $\mathbf{Z} < \mathbf{z}_1$ e outra para $\mathbf{Z} \geq \mathbf{z}_1$.

Definição 3.36. Seja $D_< = \{i : \mathbf{Z}_i < \mathbf{z}_1\}$ o conjunto de unidades amostrais em que $\mathbf{Z}_i < \mathbf{z}_1$, $\widehat{\mathbb{E}}_<[Y|\mathbf{Z}]$ e $\widehat{f}_<(X = 1|\mathbf{Z})$ regressões ajustadas utilizando apenas dados em $D_<$ e $\widehat{\mathbb{E}}_\geq[Y|\mathbf{Z}]$ e $\widehat{f}_\geq(X = 1|\mathbf{Z})$ ajustadas em $D_<^c$. O estimador RDD para $CACE(\mathbf{z}_1)$ é

$$\widehat{CACE}(\mathbf{z}_1) := \frac{\widehat{\mathbb{E}}_\geq[Y|\mathbf{z}_1] - \widehat{\mathbb{E}}_<[Y|\mathbf{z}_1]}{\widehat{f}_\geq(X = 1|\mathbf{Z}) - \widehat{f}_<(X = 1|\mathbf{Z})}.$$

Em particular, se sabemos a priori que $f(X = 1|\mathbf{z}) = 1$ para $\mathbf{z} \geq \mathbf{z}_1$ e $f(X = 1|\mathbf{z}) = 0$ para $\mathbf{z} < \mathbf{z}_1$, então

$$\widehat{CACE}(\mathbf{z}_1) := \widehat{\mathbb{E}}_\geq[Y|\mathbf{z}_1] - \widehat{\mathbb{E}}_<[Y|\mathbf{z}_1]$$

O exemplo a seguir ilustra a implementação de RDD quando $X \equiv \mathbb{I}(\mathbf{Z} \geq \mathbf{z}_1)$ utilizando tanto regressão linear quanto regressão de Kernel de Nadaraya-Watson.

Exemplo 3.37. Considere que Z_i satisfaz o critério backdoor para estimar o efeito causal de X em Y. Além disso, $Z_i \sim N(0, 1)$, $X_i \equiv \mathbb{I}(Z_i \geq 0)$ e $Y_i|X_i, Z_i \sim N(50(X_i + 1)(Z_i + 1), 1)$. Podemos simular os dados da seguinte forma:

```
n <- 1000
Z <- rnorm(n)
X <- Z >= 0
Y <- rnorm(n, 50*(X+1)*(Z+1))
data <- tibble(X, Y, Z)
plot(Z, Y)
```

Como estamos simulando os dados, podemos calcular $CACE(0)$:

$$\begin{aligned}
CACE(0) &= \lim_{\mathbf{z}\downarrow 0} \mathbb{E}[Y|\mathbf{Z} = \mathbf{z}] - \lim_{\mathbf{z}\uparrow 0} \mathbb{E}[Y|\mathbf{Z} = \mathbf{z}] && \text{Teorema 3.35} \\
&= \lim_{\mathbf{z}\downarrow 0} \mathbb{E}[Y|X = 1, \mathbf{Z} = \mathbf{z}] - \lim_{\mathbf{z}\uparrow 0} \mathbb{E}[Y|X = 0, \mathbf{Z} = \mathbf{z}] \\
&= \lim_{\mathbf{z}\downarrow 0} 50(1 + 1)(\mathbf{z} + 1) - \lim_{\mathbf{z}\uparrow 0} 50(0 + 1)(\mathbf{z} + 1) = 50
\end{aligned}$$

O código abaixo estima $CACE(0)$ usando regressão linear:

```
regs = data %>%
  mutate(Z1 = (Z >= 0)) %>%
  group_by(Z1) %>%
  summarise(
```

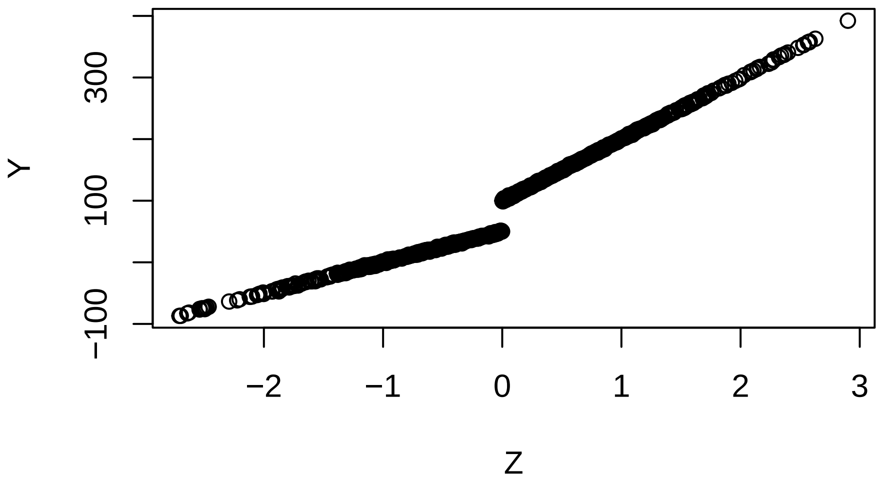

Figura 3.7.: Exemplo em que Z satisfaz o critério backdoor para medir o efeito causal de X em Y e X = I(Z > 0). Como resultado da descontinuidade da propensidade de X em Z = 0, há uma descontinuidade na regressão de Y em Z no ponto Z=0.

```
    intercepto = lm(Y ~ Z)$coefficients[1],
    coef_angular = lm(Y ~ Z)$coefficients[2]
  )
regs

## # A tibble: 2 x 3
##   Z1    intercepto coef_angular
##   <lgl>      <dbl>        <dbl>
## 1 FALSE       50.1         50.1
## 2 TRUE        99.9        100.

est_cace = 1*regs[2, 2] + 0*regs[2, 3] -
  1*regs[1, 2] + 0*regs[1, 3]
round(as.numeric(est_cace), 2)

## [1] 49.81
```

Similarmente, podemos estimar $CACE(0)$ usando regressão por kernel de Nadaraya-Watson:

```
library(np)
options(np.messages = FALSE)
nw_reg <- function(data, valor)
{
  bw <- npregbw(xdat = data$Z, ydat = data$Y)$bw
  npksum(txdat= data$Z, exdat = valor, tydat = data$Y, bws = bw)$ksum/
```

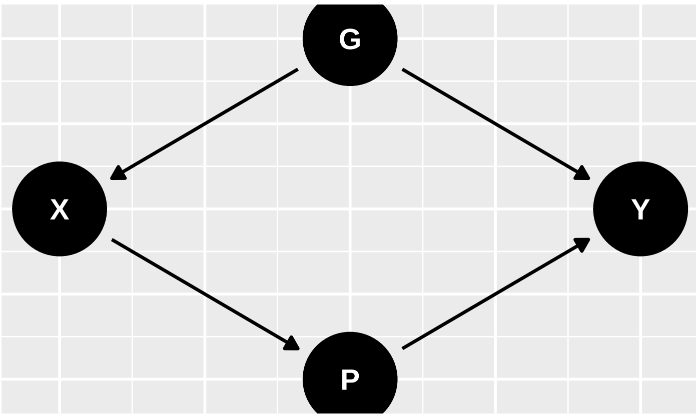

Figura 3.8.: .

```
    npksum(txdat = data$Z, exdat = valor, bws = bw)$ksum
}

reg_baixo <- data %>%
  filter(Z < 0) %>%
  nw_reg(0)
reg_cima <- data %>%
  filter(Z >= 0) %>%
  nw_reg(0)
est_cace <- reg_cima - reg_baixo
round(est_cace, 2)

## [1] 51.31
```

3.2.5. Exercícios

Exercício 3.38. Crie um exemplo em que, ao contrário do Exemplo 3.37, $\mathbb{E}[Y|X = 1, \mathbf{Z}]$ não é linear em $\mathbf{Z}$. Compare as estimativas de $CACE$ usando a regressão linear e algum método de regressão não-paramétrica.

3.3. Controlando mediadores (critério frontdoor)

Há casos em que não existem variáveis observadas que satisfazem o critério backdoor. Por exemplo, considere o grafo causal na fig. 3.8 (Glymour et al., 2016). Neste grafo, estamos interessados em compreender o efeito causal do fumo (X) sobre a incidência de câncer (Y). Além disso, fatores genéticos não observáveis (G) são um potencial confundidor, uma vez que podem ter influência tanto sobre o fumo quanto sobre a incidência de câncer. Assim, como G não é observado, não é possível implementar os métodos de estimação vistos na última seção. Apesar desta dificuldade, ainda é possível medir o efeito causal de X em Y na fig. 3.8.

Para tal, primeiramente observe que é possível estimar o efeito causal de X em P e de P em Y. Para medir o efeito causal de X em P, note que $\emptyset$ satisfaz o critério backdoor. Isso ocorre pois Y é um colisor em $X \leftarrow G \rightarrow$

$Y \leftarrow P$. Além disso, como $X = Pa(P)$, decorre do Lema 3.15 que X satisfaz o critério backdoor para medir o efeito causal de P em Y. Das duas últimas conclusões decorre do Teorema 3.16 que $f(P|do(X)) = f(P|X)$ e que $f(Y|do(P)) = \int f(Y|P,X)f(X)dX$.

A seguir, o critério frontdoor consiste em observar que P está no único caminho direcionado de X a Y, $X \rightarrow P \rightarrow Y$. Assim, é possível provar a identificação causal

$$\begin{aligned} f(Y|do(X)) &= \int f(P|do(X))f(Y|do(P))dP \\ &= \int f(P|do(X)) \int f(Y|P,X)f(X)dX. \end{aligned}$$

O critério frontdoor é formalizado a seguir:

Definição 3.39. $\mathbf{W}$ satisfaz o critério frontdoor para medir o efeito causal de X em Y se:

1. para todo caminho direcionado de X em Y, C, existe $C_i \in \mathbf{W}$ e, para todo $W \in \mathbf{W}$, existe caminho direcionado de X em Y, C, e i tal que $C_i = W$.
2. $\emptyset$ satisfaz o item 2 do critério backdoor (Definição 3.9) para medir o efeito causal de X em $\mathbf{W}$.
3. X satisfaz o item 2 do critério backdoor (Definição 3.9) para medir o efeito causal de $\mathbf{W}$ em Y.

A Definição 3.39 elenca todos os itens que utilizamos na análise da fig. 3.8. O primeiro item do critério identifica que $\mathbf{W}$ deve interceptar todos os caminhos direcionados de X a Y. Isto é, $\mathbf{W}$ capturar a informação de todos os mediadores de X a Y. O segundo e terceiro itens estabelecem as condições para que seja possível aplicar o critério backdoor para identificar $f(\mathbf{W}|do(X))$ e $f(Y|do(\mathbf{W}))$.

Identificação causal

O critério frontdoor possibilita a identificação do efeito causal de X em Y:

Teorema 3.40. *Se $\mathbf{W}$ satisfaz o critério frontdoor para medir o efeito causal de X em Y, então*

$$f(Y|do(X=x)) = \int f(\mathbf{W}|x) \int f(Y|\mathbf{W},X)f(X)dXd\mathbf{W}$$

Teorema 3.41. *Se $\mathbf{W}$ satisfaz o critério frontdoor para estimar o efeito causal de X em Y, então*

$$\mathbb{E}[Y|do(X=x)] = \mathbb{E}\left[\frac{Y \cdot f(W|x)}{f(W|X)}\right]$$

Estimação pelo critério frontdoor

A estimação é um tema menos desenvolvido ao aplicar o critério frontdoor. Alguns estimadores não-paramétricos são apresentados em Tchetgen and Shpitser (2012). A seguir, desenvolvemos um estimador não-paramétrico mais simples inspirado na estratégia de IPW.

Definição 3.42. Considere que $\mathbf{W}$ satisfaz o critério frontdoor para medir o efeito causal de X em Y e que $\widehat{f}(\mathbf{W}|X)$ é um estimador de $f(\mathbf{W}|X)$. Um estimador do tipo IPW para $\mathbb{E}[Y|do(X=x)]$ é dado por

$$\widehat{\mathbb{E}}_f[Y|do(X=x)] := n^{-1}\sum_{i=1}^{n} \frac{Y_i\widehat{f}(\mathbf{W}_i|x)}{\widehat{f}(\mathbf{W}_i|X_i)}.$$

Para provar o Teorema 3.40 utilizamos o *do calculus*, que é discutido na Seção 3.4.

3.4. Do-calculus

O *do calculus* consiste em um conjunto de regras para alterar densidade envolvend o operador "do". Por exemplo, o do calculus explica como remover o operador do, trocá-lo pelo condicionamento simples, ou remover algum condicionamento simples. Para apresentar o do calculus, é necessário primeiramente definir algumas modificações sobre o grafo causal.

Definição 3.43. Seja $(\mathcal{G}, f)$ um SCM tal que $\mathcal{G} = (\mathcal{V}, \mathcal{E})$:

$$\begin{aligned}
\mathcal{G}(\bar{\mathbb{V}}) &:= (\mathcal{V}, \{E \in \mathcal{E} : E_2 \notin \mathbb{V}\}) \\
\mathcal{G}(\bar{\mathbb{V}}_1, \underline{\mathbb{V}}_2) &:= (\mathcal{V}, \{E \in \mathcal{E} : E_2 \notin \mathbb{V}_1 \text{ e } E_1 \notin \mathbb{V}_2\}) \\
\mathcal{G}(\bar{\mathbb{V}}_1, \mathbb{V}_2^+) &= (\mathcal{V} \cup \{I_V : V \in \mathbb{V}_2\}, \{E \in \mathcal{E} : E_2 \notin \mathbb{V}_1\} \cup \{(I_V, V) : V \in \mathbb{V}_2\})
\end{aligned}$$

Isto é, $\mathcal{G}(\bar{\mathbb{V}})$ é o grafo obtido retirando de $\mathcal{G}$ as arestas que apontam para $\mathbb{V}$, $\mathcal{G}(\bar{\mathbb{V}}_1, \underline{\mathbb{V}}_2)$ é o grafo obtido retirando de $\mathcal{G}$ todas as arestas que apontam para $\mathbb{V}_1$ ou que saem de $\mathbb{V}_2$, e $\mathcal{G}(\bar{\mathbb{V}}_1, \mathbb{V}_2^+)$ é o grafo obtido adicionando a $\mathcal{G}$ um novo vértice I_V e uma aresta $I_V \rightarrow V$, para todo $V \in \mathbb{V}_2$, e retirando todas as arestas que apontam para $\mathbb{V}_1$.

Com base na Definição 3.43, é possível apresentar o do calculus:

Teorema 3.44. *Seja $(\mathcal{G}, f)$ um SCM e $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{W}$ e $\mathbf{Z}$ conjuntos de vértices disjuntos:*

1. *Se $\mathbf{Y} \perp^d \mathbf{Z} | \mathbf{X} \cup \mathbf{W}$ em $\mathcal{G}(\bar{\mathbf{X}})$, então $f(\mathbf{Y}|do(\mathbf{X}), \mathbf{Z}, \mathbf{W}) = f(\mathbf{Y}|do(\mathbf{X}), \mathbf{W})$.*

2. *Se $\mathbf{Y} \perp^d \mathbf{W} | \mathbf{Z} \cup \mathbf{X}$ em $\mathcal{G}(\bar{\mathbf{X}}, \underline{\mathbf{W}})$, então $f(\mathbf{Y}|do(\mathbf{X}), do(\mathbf{W}), \mathbf{Z}) = f(\mathbf{Y}|do(\mathbf{X}), \mathbf{W}, \mathbf{Z})$.*

3. *Se $\mathbf{Y} \perp^d I_{\mathbf{X}} | \mathbf{Z} \cup \mathbf{W}$ em $\mathcal{G}(\bar{\mathbf{W}}, \mathbf{X}^+)$, então $f(Y|do(\mathbf{W}), do(\mathbf{X}), \mathbf{Z}) = f(Y|do(\mathbf{W}), \mathbf{Z})$.*

O seguinte lema mostra como o do calculus generaliza certos aspectos do critério backdoor:

Lema 3.45. *X satisfaz o item 2 do critério backdoor para medir o efeito causal de $\mathbf{W}$ em Y se e somente se $Y \perp^d \mathbf{W} | X$ em $\mathcal{G}(\underline{\mathbf{W}})$.*

Utilizando o do calculus, é possível obter todas as relações de identificação que são válidas supondo apenas que f é compatível com o grafo causal (Shpitser and Pearl, 2006; 2008). Contudo, às vezes é razoável fazer mais suposições. Discutiremos este tipo de situação no próximo capítulo.

3.4.1. Exercícios

Exercício 3.46 (Glymour et al. (2016)[p.48])**.** Considere o modelo estrutural causal em fig. 3.9.

(a) Para cada um dos pares de variáveis a seguir, determine um conjunto de outras variáveis que as d-separa: (Z_1, W), (Z_1, Z_2), (Z_1, Y), (Z_3, W), e (X, Y).

(b) Para cada par de variáveis no item anterior, determine se elas são d-separadas dado todas as demais variáveis.

(c) Determine conjuntos de variáveis que satisfazem, respectivamente, o "backdoor criterion" e o "frontdoor criterion" para estimar o efeito causal de X em Y.

(d) Considere que para cada variável, V, temos que $V \equiv \beta_V \cdot Pa(V) + \epsilon_V$, onde os ϵ são i.i.d. e normais padrão e β_V são vetores conhecidos. Isto é, a distribuição de cada variável é determinada através de uma regressão linear simples em seus pais. Determine $f(Y|do(X = x))$ utilizando a fórmula do ajuste nos 2 casos abordados no item anterior.

$$
\begin{tikzcd}
Z_1 \arrow[dr] \arrow[dd] & & Z_2 \arrow[dl] \arrow[dd] \\
& Z_3 \arrow[dl] \arrow[dr] & \\
X \arrow[r] \arrow[dr] & W_1 \arrow[r] & Y \\
& W_2 \arrow[ur] &
\end{tikzcd}
$$

Figura 3.9.: Modelo estrutural causal do Exercício 3.46

Exercício 3.47. Considere que $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ é um grafo causal e $\mathbf{X}, \mathbf{W}, \mathbf{Y} \subseteq \mathcal{V}$. Além disso, para todo caminho, $C = (C_1, \ldots, C_n)$, com $C_1 = X \in \mathbf{X}$, $C_n = Y \in \mathbf{Y}$, e com $X \rightarrow C_2$, C está bloqueado dado $\mathbf{W}$. Prove que $f(\mathbf{y}|do(\mathbf{X})) = \int f(\mathbf{y}|\mathbf{w})f(\mathbf{w}|do(\mathbf{X}))d\mathbf{w}$ e $\mathbb{E}[Y|do(\mathbf{X})] = \mathbb{E}[\mathbb{E}[Y|\mathbf{W}]|do(\mathbf{X})]$.

Exercício 3.48. Prove que se $\mathbf{W}$ satisfaz o critério frontdoor para medir o efeito causal de X em Y, então $f(\mathbf{W}|do(X)) = f(\mathbf{W}|X)$ e $f(Y|do(\mathbf{W})) = \int f(Y|\mathbf{W}, X = x^*)f(X = x^*)dx^*$.

Exercício 3.49. Prove o Lema 3.45.

4. Resultados potenciais

No capítulo passado, vimos que $f(y|do(x))$ nos permite entender o comportamento de Y em um cenário distinto dos dados observados. Por exemplo, se X é a indicadora de um tratamento e Y é a indicadora de cura, então $f(y|do(X = 1))$ nos permite entender a proporção de cura em um cenário hipotético em que administramos o tratamento a todos os indivíduos. Esta distribuição nos permite investigar questões causais que não eram acessíveis usando apenas a distribuição observacional, $f(y, x)$.

Contudo, algumas perguntas causais não são respondidas utilizando apenas os mecanismos desenvolvidos no capítulo 3. Por exemplo, qual a probabilidade de que um indivíduo se cure quando rece o tratamento e não se cure quando não o recebe. Quando tentamos traduzir esta questão, notamos que partes dela envolvem $Y = 1$ e $do(X = 1)$ e outras partes envolvem $Y = 0$ e $do(X = 0)$. Se tentarmos ingenuamente traduzir a questão, poderíamos obter algo como $\mathbb{P}(Y = 1, Y = 0|do(X = 1), do(X = 0))$. Contudo, a probabilidade acima não responde à pergunta colocada. Em primeiro lugar, não está definido fazermos as intervenções $do(X = 1)$ e $do(X = 0)$ na mesma unidade amostral. Além disso, mesmo que a probabilidade estivesse definida, é impossível que o mesmo Y assuma tanto o valor 1 quanto 0. Isto é, $\mathbb{P}(Y = 1, Y = 0| \ldots) = 0$.

A última constatação nos revela que o modelo no capítulo 3 não tem variáveis suficientes para traduzir a perguntada levantada. Se imaginamos que é possível que um indivíduo se cure ao receber o tratamento e não se cure quando não o recebe, isto ocorre pois as ocorrências de cura em cada cenário hipotético não são logicamente equivalentes. Em outras palavras, é como se houvessem *resultados potenciais*[1], Y_1 e Y_0, para indicar a ocorrência de cura em cada cenário considerado. Com uso destas variáveis, poderíamos escrever $\mathbb{P}(Y_1 = 1, Y_0 = 0)$.

O objetivo desta seção é incluir este tipo de variável de forma a preservar as ferramentas desenvolvidas no capítulo 3. Neste quesito, a maior dificuldade será estabelecer a distribuição conjunta entre os resultados potenciais. Para tal, será útil relembrar um lema fundamental em simulação:

Lema 4.1. *Considere que $F(v|Pa(V))$ é uma função de densidade acumulada condicional arbitrária e $U \sim U(0, 1)$. Se definirmos, $V \equiv F^{-1}(U|Pa(V))$, então $V|Pa(V) \sim F$.*

O Lema 4.1 traz várias interpretações que nos serão úteis. A primeira interpretação, de caráter técnico, é que podemos simular de qualquer distribuição multivariada utilizando apenas variáveis i.i.d. e funções determinísticas. Em particular, podemos reescrever um SCM de tal forma que cada vértice, V, seja função determinística de seus pais e uma variável de ruído, U_V. Esta abordagem, que está ligada a modelos de equações estruturais, é apresentada nas Definições 4.2 e 4.3.

Definição 4.2. Seja $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ um grafo causal. O grafo causal estrutural, $\mathcal{G}^+ = (\mathcal{V}^+, \mathcal{E}^+)$, é tal que $\mathcal{V}^+ = \mathcal{V} \cup (U_V)_{V \in \mathcal{V}}$ e $\mathcal{E}^+ = \mathcal{E} \cup \{(U_V, V) : V \in \mathcal{V}\}$. Isto é, para cada $V \in \mathcal{V}$, $\mathcal{G}^+$ adiciona uma nova variável U_V e uma aresta de U_V a V.

Definição 4.3. Seja $(\mathcal{G}, f)$ um SCM. Um SCM em equações estruturais para $(\mathcal{G}, f)$, $(\mathcal{G}^+, f^+)$, é tal que $\mathcal{G}^+$ é o grafo causal estrutural de $\mathcal{G}$, $(U_V)_{V \in \mathcal{V}}$ são independentes segundo f^+ e, para cada $V \in \mathcal{V}$, existe uma função determinística, $g_V : U_V \times Pa(V) \rightarrow \Re$, tal que $f^+(V|U_V, Pa(V)) = \mathbb{I}(V = g_V(U_V, Pa(V)))$ e $f^+(\mathcal{V}) = f(\mathcal{V})$.

[1]Esta é uma tradução livre da expressão “potential outcomes” usada em inglês.

O Exemplo 4.4 ilustra uma forma de obter um SCM em equações estruturais a partir de um SCM com dois vértices.

Exemplo 4.4. Considere que $X \to Y$, $X \sim \text{Exp}(1)$ e $Y|X \sim \text{Exp}(X)$. Neste caso, o grafo estrutural causal é dado por $U_X \to X \to Y \leftarrow U_Y$. Além disso, existem várias representações do SCM em equações estruturais. Uma possibilidade é definir que U_X e U_Y são i.i.d. e $U(0,1)$, $X \equiv -\log(U_X)$ e $Y \equiv -\log(U_Y)/X$.

O Lema 4.1 também permite uma interpretação de caráter mais filosófico. Podemos imaginar que toda variável em um SCM, V, é uma função determinística de seus pais e de *condições locais* não-observadas, U_V. A expressão "condições locais" indica que cada U_V é usada somente para gerar V e que as variáveis em U são independentes, isto é, não trazem informação umas sobre as outras.

A interpretação acima é usada na definição de resultados potenciais. A ideia principal é que as mesmas funções determínistas e variáveis de ruído locais são usadas para gerar todos os resultados potenciais. A única diferença é que, para cada resultado potencial, o valor das variáveis em que houve intervenção é fixado. Esta definição é compatível com a ideia de que não é possível modificar os ruídos locais por meio da intervenção. Em outras palavras, o resultado potencial é o mais próximo possível do resultado observado sob a restrição que fixamos os valores das variáveis em que houve intervenção.

Definição 4.5. Seja $(\mathcal{G}, f)$ um SCM de grafo causal $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ e $(\mathcal{G}^+, f^+)$ o seu SCM em equações estruturais. O grafo de resultados potenciais dado por intervenções em $\mathbf{X} \subseteq \mathcal{V}$, $\mathcal{G}^* = (\mathcal{V}^*, \mathcal{E}^*)$ é tal que

$$\begin{aligned}
\mathcal{V}^* &= \{W_{\mathbb{V}=\mathbf{v}} : W \in \mathcal{V}, \mathbb{V} \subseteq \mathbf{X}, \mathbf{v} \in \mathit{supp}(\mathbb{V})\} \cup \{U_W : W \in \mathcal{V}\}, \\
\mathcal{E}^* &= \{(W_{\mathbb{V}=\mathbf{v}}, Z_{\mathbb{V}=\mathbf{v}}) : \mathbb{V} \subseteq \mathbf{X}, \mathbf{v} \in \mathit{supp}(\mathbb{V}), (W, Z) \in \mathcal{E}+, Z \notin \mathbb{V}\}.
\end{aligned}$$

Para todo $W \in \mathcal{V}$, abreviamos $W_\emptyset$ por W.

Em palavras, o grafo de resultados potenciais cria uma cópia de $\mathcal{G}$ para cada possível intervenção, $\mathbb{V} = \mathbf{v}$. Além disso, adiciona-se uma aresta de U_W para cada cópia de W. Esta construção indica que as mesmas variáveis em U geram todos os resultados potenciais. Também, para cada vértice em que houve uma intervenção, $W_{\mathbb{V}=\mathbf{v}} \in \mathbb{V}_{\mathbb{V}=\mathbf{v}}$, removem-se todas as arestas que apontam para $W_{\mathbb{V}=\mathbf{v}}$. Esta remoção ocorre porque, quando realizamos uma intervenção em $\mathbb{V}$, o valor de $\mathbb{V}$ é fixado e, assim, não é gerado por suas causas em $\mathcal{G}$.

Exemplo 4.6. Considere que $(X, Y) \in \{0,1\}^2$ e o grafo causal é $X \to Y$. Vimos no Exemplo 4.4 que o grafo causal estrutural é dado por $U_X \to X \to Y \leftarrow U_Y$. Vamos construir o grafo de resultados potenciais dadas intervenções em X. Neste caso, além dos vértices U_X, U_Y, X, Y, temos também $X_{X=0}, Y_{X=0}, X_{X=1}, Y_{X=1}$. Como não há ambiguidade neste caso, podemos abreviar os últimos quatro vértices por X_0, Y_0, X_1, Y_1.

O grafo de resultados potenciais é ilustrado na fig. 4.1. O grafo causal estrutural é a reta horizontal de U_X a U_Y. Os resultados potenciais são cópias deste grafo que usam as mesmas variáveis U e em que removemos as arestas que apontam para a intervenções, X_0 e X_1.

Uma vez definido o grafo de resultados potenciais, podemos extender a distribuição do modelo de equações estruturais para este grafo. Esta extensão envolve três etapas. Primeiramente, a distribuição de U continua a mesma. Em segundo lugar, para todo vértice do grafo de resultados potenciais, $W_{\mathbb{V}=\mathbf{v}}$, em que não houve uma intervenção, este vértice é gerado pelo mesmo mecanismo que W. Isto é, $W_{\mathbb{V}=\mathbf{v}} = \mathbb{I}(g_W(U_W, Pa^*(W_{\mathbb{V}=\mathbf{v}})))$. Finalmente, se houve uma intervenção em $W_{\mathbb{V}=\mathbf{v}}$, então ela é uma variável degenerada no valor desta intervenção. Esta construção é formalizada na Definição 4.7.

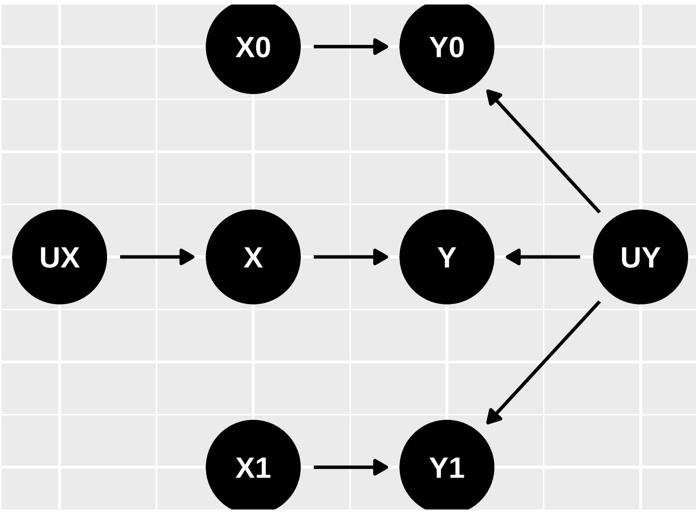

Figura 4.1.: Grafo de resultados potenciais dadas intervenções em X.

Definição 4.7. Seja $(\mathcal{G}^+, f^+)$ um SCM em equações estruturais para $(\mathcal{G}, f)$ com funções determinísticas, g. O modelo de resultados potenciais (POM)[2] dado por intervenções em $\mathbf{X}$, $(\mathcal{G}^*, f^*)$ é tal que $\mathcal{G}^*$ é o grafo de resultados potenciais dado por intervenções em $\mathbf{X}$ (Definição 4.5) e

$$f^*(U_W) = f(U_W) \qquad \text{, para todo } W \in \mathcal{V},$$
$$f^*(W_{\mathbb{V}=\mathbf{v}}|Pa^*(W_{\mathbb{V}=\mathbf{v}})) = \begin{cases} \mathbb{I}(W_{\mathbb{V}=\mathbf{v}} = \mathbf{v}_i) & \text{, se } W \equiv \mathbb{V}_i \\ \mathbb{I}(W_{\mathbb{V}=\mathbf{v}} = g_W(U_W, Pa^*(W_{\mathbb{V}=\mathbf{v}}))) & \text{, caso contrário.} \end{cases}$$

Ainda que seja uma formalização conveniente, o POM é consideravelmente mais complexo que o SCM original. Para ganhar intuição sobre o POM, alguns lemas de tradução são fundamentais.

Lema 4.8. *Para toda intervenção,* $\mathbb{V} = \mathbf{v}$, $\mathbb{P}(\mathcal{V} = \mathcal{V}_{\mathbb{V}=\mathbf{v}}|\mathbb{V} = \mathbf{v}) = 1$.

O Lema 4.8 conecta o dado observacional em $\mathcal{V}$ ao resultado potencial $\mathcal{V}_{\mathbb{V}=\mathbf{v}}$. Mais especificamente, quando observamos que $\mathbb{V} = \mathbf{v}$, então os resultados potenciais dada a intervenção $\mathbb{V} = \mathbf{v}$ são idênticos aos resultados observados. Em outras palavras, ao observamos que $\mathbb{V} = \mathbf{v}$, aprendemos que estamos justamente na hipótese de resultados potenciais em que $\mathbb{V} = \mathbf{v}$.

Lema 4.9.

$$f(\mathcal{V}_{\mathbb{V}=\mathbf{v}}) = f(\mathcal{V}|do(\mathbb{V} = \mathbf{v}))$$

Lema 4.10 (Ignorabilidade condicional)**.** *Se* $\mathbf{Z}$ *satisfaz o critério backdoor para medir o efeito causal de* X *em* Y*, então para todo* x*,* $Y_{X=x} \perp^d X|\mathbf{Z}$.

4.0.1. Exercícios

Exercício 4.11. Prove o Lema 4.1.

[2] utilizamos a sigla POM em referência ao termo em inglês "potential outcomes model"

Exercício 4.12. Mostre que no Exemplo 4.4 a distribuição de (X, Y) no SCM em equações estruturais é igual àquela no SCM original.

Bibliografia

Barrett, M. (2022). *ggdag: Analyze and Create Elegant Directed Acyclic Graphs*. R package version 0.2.7.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., and Yuan, J. (2023). *xgboost: Extreme Gradient Boosting*. R package version 1.7.3.1.

Glymour, M., Pearl, J., and Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.

Hahn, J., Todd, P., and Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209.

Lee, D. S. and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of economic literature*, 48(2):281–355.

Mauá, D. (2022). Probabilistic Graphical Models. https://www.ime.usp.br/~ddm/courses/mac6916/. [Online; accessed 22-October-2022].

Sackett, D. L. (1979). Bias in analytic research. *Journal of Chronic Diseases*, 32(1-2):51–63.

Shpitser, I. and Pearl, J. (2006). Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1219. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Shpitser, I. and Pearl, J. (2008). Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241.

Tchetgen, E. J. T. and Shpitser, I. (2012). Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of statistics*, 40(3):1816.

Textor, J., Van der Zander, B., Gilthorpe, M. S., Liśkiewicz, M., and Ellison, G. T. (2016). Robust causal inference using directed acyclic graphs: the r package 'dagitty'. *International journal of epidemiology*, 45(6):1887–1894.

A. Demonstrações

A.1. Relativas à Seção 2.2

A.1.1. Relativas ao Lema 2.36

Prova do Lema 2.36. A prova consistirá em demonstrar que, para cada i, a afirmação i decorre da afirmação $i-1$. Finalmente, a afirmação 1 decorre da afirmação 4. Os símbolos $\mathbf{X}$ e $\mathbf{x}$ referem-se a $(\mathbf{X}_1, \ldots, \mathbf{X}_d)$ e $(\mathbf{x}_1, \ldots, \mathbf{x}_d)$.

- $(1 \Longrightarrow 2)$

$$\begin{aligned} f(\mathbf{x}|\mathbf{y}) &= \prod_{j=1}^{d} f(\mathbf{x}_j|\mathbf{y}) && (1) \\ &= \prod_{j=1}^{d} h(\mathbf{x}_j, \mathbf{y}) && h(\mathbf{x}_j, \mathbf{y}) = f(\mathbf{x}_j|\mathbf{y}) \end{aligned}$$

- $(2 \Longrightarrow 3)$ Note que,

$$\begin{aligned} f(\mathbf{x}_i|\mathbf{x}_{-i}, \mathbf{y}) &= \frac{f(\mathbf{x}|\mathbf{y})}{f(\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots \mathbf{x}_d|\mathbf{y})} \\ &= \frac{f(\mathbf{x}|\mathbf{y})}{\int_{\mathbb{R}} f(\mathbf{x}|\mathbf{y}) d\mathbf{x}_i} \\ &= \frac{\prod_{j=1}^{d} h_j(\mathbf{x}_j, \mathbf{y})}{\int_{\mathbb{R}} \prod_{j=1}^{d} h_j(\mathbf{x}_j, \mathbf{y}) d\mathbf{x}_i} (2) \\ &= \frac{\prod_{j=1}^{d} h_j(\mathbf{x}_j, \mathbf{y})}{\prod_{j \neq i} h_j(\mathbf{x}_j, \mathbf{y}) \int_{\mathbb{R}} h_i(\mathbf{x}_i, \mathbf{y}) d\mathbf{x}_i} \\ &= \frac{\tilde{h}_i(\mathbf{x}_i, \mathbf{y})}{\int_{\mathbb{R}} h_i(\mathbf{x}_i, \mathbf{y}) d\mathbf{x}_i} \\ &= \frac{\prod_{j \neq i} \int_{\mathbb{R}} h_j(\mathbf{x}_j, \mathbf{y}) d\mathbf{x}_j}{\prod_{j \neq i} \int_{\mathbb{R}} h_j(\mathbf{x}_j, \mathbf{y}) d\mathbf{x}_j} \cdot \frac{h_i(\mathbf{x}_i, \mathbf{y})}{\int_{\mathbb{R}} h_i(\mathbf{x}_i, \mathbf{y}) d\mathbf{x}_i} \\ &= \frac{\int_{\mathbb{R}^{d-1}} \prod_{j=1}^{d} h_j(\mathbf{x}_j, \mathbf{y}) d\mathbf{x}_{-i}}{\int_{\mathbb{R}^d} \prod_{j=1}^{d} h_j(\mathbf{x}_j, \mathbf{y}) d\mathbf{x}} \\ &= \frac{\int_{\mathbb{R}^{d-1}} f(\mathbf{x}|\mathbf{y}) d\mathbf{x}_{-i}}{\int_{\mathbb{R}^d} f(\mathbf{x}|\mathbf{y}) d\mathbf{x}} && (2) \\ &= f(\mathbf{x}_i|\mathbf{y}) \end{aligned}$$

- $(3 \implies 4)$

$$
\begin{aligned}
f(\mathbf{x}_i|\mathbf{x}_1^{i-1},\mathbf{y}) &= \frac{f(\mathbf{x}_1^i|\mathbf{y})}{f(\mathbf{x}_1^{i-1}|\mathbf{y})} \\
&= \frac{\int_{\mathbb{R}^{d-i}} f(\mathbf{x}|\mathbf{y}) d\mathbf{x}_{i+1}^d}{f(\mathbf{x}_1^{i-1}|\mathbf{y})} \\
&= \frac{\int_{\mathbb{R}^{d-i}} f(\mathbf{x}_{-i}|\mathbf{y}) f(\mathbf{x}_i|\mathbf{x}_{-i},\mathbf{y}) d\mathbf{x}_{i+1}^d}{f(\mathbf{x}_1^{i-1}|\mathbf{y})} \\
&= \frac{f(\mathbf{x}_i|\mathbf{y}) \int_{\mathbb{R}^{d-i}} f(\mathbf{x}_{-i}|\mathbf{y}) d\mathbf{x}_{i+1}^d}{f(\mathbf{x}_1^{i-1}|\mathbf{y})} \\
&= \frac{f(\mathbf{x}_i|\mathbf{y}) f(\mathbf{x}_1^{i-1}|\mathbf{y})}{f(\mathbf{x}_1^{i-1}|\mathbf{y})} \\
&= f(\mathbf{x}_i|\mathbf{y})
\end{aligned} \tag{3}
$$

- $(4 \implies 1)$

$$
\begin{aligned}
f(\mathbf{x}|\mathbf{y}) &= \prod_{i=1}^{d} f(\mathbf{x}_i|\mathbf{x}_1^{i-1},\mathbf{y}) \\
&= \prod_{i=1}^{d} f(\mathbf{x}_i|\mathbf{y})
\end{aligned} \tag{4}
$$

□

A.1.2. Relativas ao Teorema 2.40

Lema A.1. *Seja $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ um DAG. Se $\mathcal{A} = \mathbb{V}_1 \cup \mathbb{V}_2 \cup \mathbb{V}_3$ é ancestral e $\mathbb{V}_1 \perp \mathbb{V}_2 | \mathbb{V}_3$, então, para todo f compatível com $\mathcal{G}$, $\mathbb{V}_1 \perp\!\!\!\perp^f \mathbb{V}_2 | \mathbb{V}_3$.*

Demonstração. Defina $\mathbb{V}_1^* = \{V \in \mathcal{A} : V \in \mathbb{V}_1 \text{ ou } V_1 \rightarrow V, \text{ para algum } V_1 \in \mathbb{V}_1\}$ e $\mathbb{V}_2^* = \mathcal{A} - \mathbb{V}_1^*$. Como $\mathbb{V}_1 \perp \mathbb{V}_2 | \mathbb{V}_3$, decorre de Definição 2.39 que não existe $V_1 \in \mathbb{V}_1$ e $V_2 \in \mathbb{V}_2$ tal que $V_1 \rightarrow V_2$. Portanto,

$$
\mathbb{V}_1^* \subseteq \mathbb{V}_1 \cup \mathbb{V}_3 \text{ e } \mathbb{V}_2^* \subseteq \mathbb{V}_2 \cup \mathbb{V}_3 \tag{A.1}
$$

A seguir, demonstraremos que

$$
\forall i \in \{1, 2\} \text{ e } V_i^* \in \mathbb{V}_i^* : Pa(V_i^*) \subseteq \mathbb{V}_i \cup \mathbb{V}_3 \tag{A.2}
$$

Tome $V_1^* \in \mathbb{V}_1^*$. Como $V_1^* \in \mathcal{A}$ e $\mathcal{A}$ é ancestral, decorre da Definição 2.8 que $Pa(V_1^*) \subseteq \mathcal{A}$. Assim, basta demonstrar que $Pa(V_1^*) \cap \mathbb{V}_2 = \emptyset$. Se $V_1^* \in \mathbb{V}_1$, então decorre de Definição 2.39 que não existe $V_2 \in \mathbb{V}_2$ tal que $V_2 \rightarrow V_1^*$. Caso contrário, se $V_1^* \in \mathbb{V}_3$, então existe $V_1 \in \mathbb{V}_1$ tal que $V_1 \rightarrow V_1^*$. Decorre de Definição 2.39 que não existe $V_1 \in \mathbb{V}_1$, $V_2 \in \mathbb{V}_2$ e $V_3 \in \mathbb{V}_3$ tais que V_3 é um colisor entre V_1 e V_2, isto é, $V_1 \rightarrow V_3 \leftarrow V_2$. Portanto, não existe $V_2 \in \mathbb{V}_2$ tal que $V_2 \rightarrow V_1^*$. Conclua que $Pa(V_1^*) \subseteq \mathbb{V}_1 \cup \mathbb{V}_3$.

A seguir, note que pela definição de $\mathbb{V}_1^*$, se $V \in \mathcal{A}$ é tal que existe $V_1 \in \mathbb{V}_1$ com $V_1 \rightarrow V$, então $V \in \mathbb{V}_1^*$. Portanto, como $\mathbb{V}_2^* = \mathcal{V} - \mathbb{V}_1^*$, para todo $V_2^* \in \mathbb{V}_2^*$, não existe $V_1 \in \mathbb{V}_1$ tal que $V_1 \rightarrow V_2^*$. Isto é, $Pa(V_2^*) \subseteq \mathcal{V} - \mathbb{V}_1$. Como $V_2^* \in \mathcal{A}$ e $\mathcal{A}$ é ancestral, conclua da Definição 2.8 que $Pa(V_2^*) \subseteq \mathcal{A}$. Combinando as duas últimas frases,

$Pa(V_2^*) \subseteq \mathbb{V}_2 \cup \mathbb{V}_3$.

Decorre da conclusão dos dois últimos parágrafos que eq. (A.2) está demonstrado.

$$\begin{aligned}
f(\mathbb{V}_1, \mathbb{V}_2|\mathbb{V}_3) &= \frac{f(\mathbb{V}_1, \mathbb{V}_2, \mathbb{V}_3)}{f(\mathbb{V}_3)} \\
&= \frac{\prod_{V \in \mathcal{A}} f(V|Pa(V))}{f(\mathbb{V}_3)} && \text{Lema 2.12} \\
&= \frac{\left(\prod_{V_1^* \in \mathbb{V}_1^*} f(V_1^*|Pa(V_1^*))\right)\left(\prod_{V_2^* \in \mathbb{V}_2^*} f(V_2^*|Pa(V_2^*))\right)}{f(\mathbb{V}_3)} && \mathbb{V}_1^* \text{ e } \mathbb{V}_2^* \text{ particionam } \mathcal{A} \\
&= h_1(\mathbb{V}_1, \mathbb{V}_3) h_2(\mathbb{V}_2, \mathbb{V}_3) && \text{eqs. (A.1) e (A.2)}
\end{aligned}$$

Assim, decorre do Lema 2.36 que $\mathbb{V}_1 \perp\!\!\!\perp^f \mathbb{V}_2|\mathbb{V}_3$. □

Lema A.2. *Se f é compatível com $\mathcal{G}$ e $\mathbb{V}_1 \perp \mathbb{V}_2|\mathbb{V}_3$, então $\mathbb{V}_1 \perp\!\!\!\perp^f \mathbb{V}_2|\mathbb{V}_3$.*

Demonstração. Defina $\mathcal{A} = Anc(\mathbb{V}_1 \cup \mathbb{V}_2 \cup \mathbb{V}_3)$, $\mathbb{V}_1^* = \{V \in \mathcal{A} : V$ não é d-separado de $\mathbb{V}_1|\mathbb{V}_3\}$, e $\mathbb{V}_2^* = \mathcal{A} - \mathbb{V}_1^*$. Por definição,

$$\mathbb{V}_1 \subseteq \mathbb{V}_1^* \text{ e } \mathbb{V}_2 \subseteq \mathbb{V}_2^* \tag{A.3}$$

O primeiro é provar que $\mathbb{V}_1^* \perp \mathbb{V}_2^*|\mathbb{V}_3$. Pela definição de $\mathbb{V}_2^*$, para todo $V_1 \in \mathbb{V}_1$ e $V_2^* \in \mathbb{V}_2^*$, $V_1 \perp V_2^*|\mathbb{V}_3$, isto é,

$$\mathbb{V}_1 \perp \mathbb{V}_2^*|\mathbb{V}_3 \tag{A.4}$$

Suponha por absurdo que existam $V_1^* \in \mathbb{V}_1^*$ e $V_2^* \in \mathbb{V}_2^*$ tais que V_1^* e V_2^* não são d-separados dado $\mathbb{V}_3$. Portanto, existe um caminho ativo dado $\mathbb{V}_3$, $(V_1^*, C_2, \ldots, C_{n-1}, V_2^*)$. Pela definição de $\mathbb{V}_1^*$, existe $V_1 \in \mathbb{V}_1$ e um caminho ativo dado $\mathbb{V}_3$, $(V_1, C_2^*, \ldots, C_{m-1}^*, V_1^*)$. Assim, $(V_1, C_2^*, \ldots, C_{m-1}^*, V_1^*, C_2, \ldots, C_{n-1}, V_2^*)$ é é um caminho ativo dado $\mathbb{V}_3$ de V_1 a V_2^*, uma contradição com eq. (A.4). Conclua que $\mathbb{V}_1^* \perp \mathbb{V}_2^*|\mathbb{V}_3$.

A seguir, provaremos que $\mathbb{V}_1^* \perp\!\!\!\perp^f \mathbb{V}_2^*|\mathbb{V}_3$. Como $\mathcal{A} = Anc(\mathbb{V}_1 \cup \mathbb{V}_2 \cup \mathbb{V}_3)$, decorre do Lema 2.9 que $\mathcal{A}$ é ancestral. Portanto, como $\mathcal{A} = \mathbb{V}_1^* \cup \mathbb{V}_2^* \cup \mathbb{V}_3$ e $\mathbb{V}_1^* \perp \mathbb{V}_2^*|\mathbb{V}_3$, decorre do Lema A.1 que $\mathbb{V}_1^* \perp\!\!\!\perp^f \mathbb{V}_2^*|\mathbb{V}_3$.

Como $\mathbb{V}_1^* \perp\!\!\!\perp^f \mathbb{V}_2^*|\mathbb{V}_3$, a conclusão do lema decorre do fato de que $\mathbb{V}_1 \subseteq \mathbb{V}_1^*$ e $\mathbb{V}_2 \subseteq \mathbb{V}_2^*$. □

Lema A.3. *Se $\mathbb{V}_1$ não é d-separado de $\mathbb{V}_2$ dado $\mathbb{V}_3$ segundo o DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, então existe f compatível com $\mathcal{G}$ tal que $\mathbb{V}_1$ e $\mathbb{V}_2$ são condicionalmente dependentes dado $\mathbb{V}_3$ segundo f*

Demonstração. □

Prova do Teorema 2.40. Decorre dos Lemas A.2 e A.3. □

A.2. Relativas à Seção 3.2

A.2.1. Relativas ao Teorema 3.16

Para realizar a demonstração do Teorema 3.16, consideraremos um SCM aumentado, em que existe uma variável que representa a ocorrência de uma intervenção em X. Uma consequência interessante desta construção será a de que o modelo intervencional é equivalente ao condicionamento usual no SCM aumentado.

Definição A.4. Seja $(\mathcal{G}_*, f_*)$ um SCM expandido tal que $\mathcal{G}_* = (\mathcal{V} \cup \{I_X : X \in \mathbf{X}\}, \mathcal{E}_*)$, e $\mathcal{E}_* = \mathcal{E} \cup \{(I_X \to X : X \in \mathbf{X})\}$. Isto é, $\mathcal{G}_*$ é uma cópia de $\mathcal{G}$ em que adicionamos para cada $X \in \mathbf{X}$ os vértice $I_X \in \{0, 1\}$ e arestas de I_X para X.

$\mathcal{G}^*$ admite uma interpretação intuitiva. I_X é a indicadora de que fazemos uma intervenção em X, fazendo que esta assuma o valor x. Se $I_X = 0$, não há uma intervenção e, assim, X segue a sua distribuição observacional. Se $I_X = 1$, X assume o valor x com probabilidade 1.

Finalmente, considerando $Pa(X)$ como os pais de X segundo $\mathcal{G}$, definimos que:

$$f_*(X|Pa(X), I_X) = \begin{cases} f(X|Pa(X)) & \text{, se } I_X = 0\text{, e} \\ \mathbb{I}(X = x) & \text{, caso contrário.} \end{cases}$$

Lema A.5. *Se $(\mathcal{G}_*, f_*)$ é tal qual em Definição A.4, então:*

$$f(\mathcal{V}|do(X = x)) = f_*(\mathcal{V}|I_{\mathbf{X}} = 1)$$

Demonstração.

$$\begin{aligned} f_*(\mathcal{V}|I_{\mathbf{X}} = 1) &= \frac{f_*(\mathcal{V}, I_{\mathbf{X}} = 1)}{f(I_{\mathbf{X}} = 1)} \\ &= \frac{f(I_{\mathbf{X}} = 1) \prod_{X \in \mathbf{X}} \mathbb{I}(X = x) \prod_{V \notin \mathbf{X}} f(V|Pa(V))}{f(I = 1)} && \text{Definição A.4} \\ &= \prod_{X \in \mathbf{X}} \mathbb{I}(X = x) \cdot \prod_{V \notin \mathbb{V}_1} f(V|Pa(V)) \\ &= f(\mathcal{V}|do(\mathbf{X} = \mathbf{x})) && \text{Definição 3.1} \end{aligned}$$

□

Lema A.6. *Se $(\mathcal{G}_*, f_*)$ é tal qual em Definição A.4, então:*

$$f_*(\mathcal{V}|I_{\mathbf{X}} = 0) = f(\mathcal{V}).$$

Demonstração.

$$\begin{aligned} f_*(\mathcal{V}|I_{\mathbf{X}} = 0) &= \frac{f_*(\mathcal{V}, I_{\mathbf{X}} = 0)}{f_*(I_{\mathbf{X}} = 0)} \\ &= \frac{f_*(I_{\mathbf{X}} = 0) \prod_{X \in \mathbf{X}} f_*(X|Pa(X), I_X = 0) \prod_{V \notin \mathbf{X}} f(V|Pa(V))}{f_*(I = 0)} && \text{Definição A.4} \\ &= \prod_{X \in \mathbf{X}} f(X|Pa(X)) \prod_{V \notin \mathbf{X}} f(V|Pa(V)) && \text{Definição A.4} \\ &= \prod_{V \in \mathcal{V}} f(V|Pa(V)) \\ &= f(\mathcal{V}) && \text{Definição 2.10} \end{aligned}$$

□

Lema A.7. *Se $(\mathcal{G}_*, f_*)$ é tal qual em Definição A.4 e $\mathbf{Z}$ satisfaz o segundo item do critério backdoor para medir o efeito causal de X em Y, então $I \perp^d Y|X, \mathbf{Z}$.*

Demonstração. Tome um caminho arbitrário de I em Y, $C = (I, C_2, \ldots, C_{n-1}, Y)$. Por definição de I, $C_2 = X$ e $I \rightarrow X$. Se $X \rightarrow C_3$, então X não é um colisor em C e C está bloqueado dado X e $\mathbf{Z}$. Se $X \leftarrow C_3$, então $(X, C_3, \ldots, C_{n-1}, Y)$ está bloqueado dado $\mathbf{Z}$, uma vez que $\mathbf{Z}$ satisfaz o segundo item do critério backdoor. Conclua que C está bloqueado dado X e $\mathbf{Z}$. $\square$

Lema A.8. *Se* $\mathbf{Z}$ *satisfaz o segundo item do critério backdoor para medir o efeito causal de* X *em* Y*, então*

$$f(y|do(x), \mathbf{z}) = f(y|x, \mathbf{z}).$$

Demonstração.

$$\begin{aligned}
f(y|do(x), \mathbf{z}) &= f_*(y|I = 1, \mathbf{z}) && \text{Lema A.5} \\
&= \int f_*(y, X|I = 1, \mathbf{z})dX \\
&= \int f_*(X|I = 1, \mathbf{z})f_*(y|X, I = 1, \mathbf{z})dX \\
&= \int \mathbb{I}(X = x)f_*(y|X, I = 1, \mathbf{z})dX \\
&= f_*(y|x, I = 1, \mathbf{z}) \\
&= f_*(y|x, I = 0, \mathbf{z}) && \text{Lema A.7} \\
&= f(y|x, \mathbf{z}) && \text{Lema A.6}
\end{aligned}$$

$\square$

Lema A.9. *Se* $(\mathcal{G}_*, f_*)$ *é tal qual em Definição A.4 e* $\mathbf{X} \notin Anc(\mathbf{Z})$*, então:*

$$f_*(\mathbf{z}) = f(\mathbf{z})$$

Demonstração. Seja $\mathbf{Z}_* = Anc(\mathbf{Z})$ e $\mathbb{V} = \mathcal{V} - (\{X\} \cup \mathbf{Z}_*)$. Como $X \notin \mathbf{Z}_*$, decorre da Definição A.4 que $I \notin \mathbf{Z}_*$. Portanto,

$$\begin{aligned}
f_*(\mathbf{z}_*) &= \int f_*(\mathbf{z}_*, I, X, \mathbf{v})d(I, X, \mathbf{v}) \\
&= \int \left(\prod_{z \in \mathbf{z}_*} f(z|Pa(z))\right)\left(f_*(I)f_*(X|I, Pa(X))\prod_{v \in \mathbf{v}} f(v|Pa(v))\right)d(I, X, \mathbf{v}) && \text{Definição A.4} \\
&= \left(\prod_{z \in \mathbf{z}_*} f(z|Pa(z))\right)\int \left(f_*(I)f_*(X|I, Pa(X))\prod_{v \in \mathbf{v}} f(v|Pa(v))\right)d(I, X, \mathbf{v}) && \mathbf{Z}_* \cap (\mathbb{V} \cup \{I, X\}) = \emptyset \\
&\propto \prod_{z \in \mathbf{z}_*} f(z|Pa(z)) && \mathbf{Z}_* \cap (\mathbb{V} \cup \{I, X\}) = \emptyset \\
&= f(\mathbf{z}_*) && \text{Exercício 2.24}
\end{aligned}$$

Assim, decorre da Lei da Probabilidade Total que $f_*(\mathbf{z}) = f(\mathbf{z})$. $\square$

Lema A.10. *Se* $(\mathcal{G}_*, f_*)$ *é tal qual em Definição A.4 e* $\mathbf{Z}$ *satisfaz o critério backdoor para medir o efeito causal de* X *em* Y*, então* $I \perp^d \mathbf{Z}$*.*

Demonstração. Tome arbitrariamente um $Z \in \mathbf{Z}$ e um caminho de I em Z, $C = (I, C_2, \ldots, C_{n-1}, Z)$. Por definição de I, $C_2 = X$ e $I \to X$. Suponha por absurdo que C não tem colisor. Como, $I \to X$, decorre que $C = I \to X \to \ldots \to C_{n-1} \to Z$. Assim, Z é um descendente de X, uma contradição com o critério backdoor (Definição 3.9). Conclua que C tem um colisor. Assim, C está marginalmente bloqueado (Definição 2.38). □

Lema A.11. *Se $\mathbf{Z}$ satisfaz o critério backdoor para medir o efeito causal de X em Y, então $f(\mathbf{z}|do(x)) = f(\mathbf{z})$.*

Demonstração.

$$\begin{aligned} f(\mathbf{z}|do(x)) &= f_*(\mathbf{z}|I=1) && \text{Lema A.5} \\ &= f_*(\mathbf{z}) && \text{Lema A.10} \\ &= f(\mathbf{z}) && \text{Lema A.9} \end{aligned}$$

□

Prova do Teorema 3.16. Decorre diretamente dos Lemas A.8 e A.11. □

Prova do Corolário 3.17.

$$\begin{aligned} f(y|do(X=x)) &= \int f(y, \mathbf{z}|do(X=x))d\mathbf{z} \\ &= \int f(\mathbf{z}|do(X=x))f(y|do(X=x), \mathbf{z}) \\ &= \int f(\mathbf{z})f(y|x, \mathbf{z}) && \text{Teorema 3.16} \end{aligned}$$

□

A.2.2. Relativas aos Teoremas 3.18 e 3.19

Prova do Teorema 3.18.

$$\begin{aligned} \mathbb{E}[g(Y)|do(X=x), \mathbf{Z}] &= \int g(y)f(y|do(x), \mathbf{Z})dy && \text{Definição 3.2} \\ &= \int g(y)f(y|x, \mathbf{Z})dy && \text{Teorema 3.16} \\ &= \mathbb{E}[g(Y)|X=x, \mathbf{Z}] && \text{(A.5)} \\ \mathbb{E}[g(Y)|do(X=x)] &= \mathbb{E}[\mathbb{E}[g(Y)|do(X=x), \mathbf{Z}]] \\ &= \mathbb{E}[\mathbb{E}[g(Y)|X=x, \mathbf{Z}]] && \text{eq. (A.5)} \end{aligned}$$

□

Prova do Teorema 3.19.

$$\begin{aligned}
\mathbb{E}[g(Y)|do(x),\mathbf{Z}] &= \int g(y)f(y|do(x),\mathbf{Z})dy && \text{Definição 3.2}\\
&= \int g(y)f(y|x,\mathbf{Z})dy && \text{Teorema 3.16}\\
&= \int \frac{g(y)f(y,x|\mathbf{Z})}{f(x|\mathbf{Z})}dy\\
&= \int \frac{g(y)\mathbb{I}(x_*=x)f(y,x_*|\mathbf{Z})}{f(x|\mathbf{Z})}d(x_*,y)\\
&= \mathbb{E}\left[\frac{g(Y)\mathbb{I}(X=x)}{f(x|\mathbf{Z})}|\mathbf{Z}\right]\\
&= \frac{\mathbb{E}[g(Y)\mathbb{I}(X=x)|\mathbf{Z}]}{f(x|\mathbf{Z})}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[Y|do(x)] &= \mathbb{E}[\mathbb{E}[Y|do(X),\mathbf{Z}]]\\
&= \mathbb{E}\left[\frac{\mathbb{E}[g(Y)\mathbb{I}(X=x)|\mathbf{Z}]}{f(x|\mathbf{Z})}\right] && \text{Teorema 3.19}\\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{g(Y)\mathbb{I}(X=x)}{f(x|\mathbf{Z})}|\mathbf{Z}\right]\right]\\
&= \mathbb{E}\left[\frac{g(Y)\mathbb{I}(X=x)}{f(x|\mathbf{Z})}\right]
\end{aligned}$$

□

A.2.3. Relativas ao Teorema 3.23

Lema A.12. *Se $(W_n)_{n\in\mathbb{N}}$ é uma sequência de variáveis aleatórias tais que $\mathbb{E}[|W_n|]=o(1)$, então $W_n \xrightarrow{\mathbb{P}} 0$.*

Demonstração.

$$\begin{aligned}
\mathbb{P}(|W_n|>\epsilon) &\leq \frac{\mathbb{E}[|W_n|]}{\epsilon} && \text{Markov}\\
&= o(1)
\end{aligned}$$

□

Prova do Teorema 3.23. Como $\mathbb{E}[|\mu(x,\mathbf{Z})|]<\infty$, pela Lei dos Grandes Números,

$$\frac{\sum_{i=1}^n \mu(x,\mathbf{Z}_i)}{n} \xrightarrow{\mathbb{P}} \mathbb{E}[\mu(x,\mathbf{Z})]$$

Portanto, pelo Teorema 3.18, é suficiente provar que $\widehat{\mathbb{E}}_1[Y|do(X=x)] - \frac{\sum_{i=1}^n \mu(x,\mathbf{Z}_i)}{n} \xrightarrow{\mathbb{P}} 0$. Usando o Lema A.12, é suficiente provar que $\mathbb{E}\left[\left|\widehat{\mathbb{E}}_1[Y|do(X=x)] - \frac{\sum_{i=1}^{n}\mu(x,\mathbf{Z}_i)}{n}\right|\right] = o(1)$.

$$\begin{aligned}
\mathbb{E}\left[\left|\widehat{\mathbb{E}}_1[Y|do(X=x)] - \frac{\sum_{i=1}^{n}\mu(x,\mathbf{Z}_i)}{n}\right|\right] &= \mathbb{E}\left[\left|\frac{\sum_{i=1}^{n}(\hat{\mu}(x,\mathbf{Z}_i)-\mu(x,\mathbf{Z}_i))}{n}\right|\right] \\
&\leq n^{-1}\sum_{i=1}^{n}\mathbb{E}\left[|\hat{\mu}(x,\mathbf{Z}_i)-\mu(x,\mathbf{Z}_i)|\right] \\
&= \mathbb{E}\left[|\hat{\mu}(x,\mathbf{Z})-\mu(x,\mathbf{Z})|\right] && \text{Definição 3.20} \\
&= o(1)]
\end{aligned}$$

□

A.2.4. Relativas ao Teorema 3.26

Prova do Teorema 3.26. Pela Lei dos Grandes números, $n^{-1}\sum_{i=1}^{n}\frac{Y_i\mathbb{I}(X_i=x)}{f(x|\mathbf{Z}_i)} \xrightarrow{\mathbb{P}} \mathbb{E}\left[\frac{Y\mathbb{I}(X=x)}{f(x|\mathbf{Z})}\right]$. Como pelo Teorema 3.19 temos que $\mathbb{E}\left[\frac{Y\mathbb{I}(X=x)}{f(x|\mathbf{Z})}\right] = \mathbb{E}[Y|do(X=x)]$, usando o Lema A.12 é suficiente provar que

$$\mathbb{E}\left[\left|n^{-1}\sum_{i=1}^{n}\frac{Y_i\mathbb{I}(X_i=x)}{\widehat{f}(x|\mathbf{Z}_i)} - n^{-1}\sum_{i=1}^{n}\frac{Y_i\mathbb{I}(X_i=x)}{f(x|\mathbf{Z}_i)}\right|\right] = o(1).$$

$$\begin{aligned}
&\mathbb{E}\left[\left|n^{-1}\sum_{i=1}^{n}\frac{Y_i\mathbb{I}(X_i=x)}{\widehat{f}(x|\mathbf{Z}_i)} - n^{-1}\sum_{i=1}^{n}\frac{Y_i\mathbb{I}(X_i=x)}{f(x|\mathbf{Z}_i)}\right|\right] \\
\leq& n^{-1}\sum_{i=1}^{n}\mathbb{E}\left[\left|\frac{Y_i\mathbb{I}(X_i=x)}{\widehat{f}(x|\mathbf{Z}_i)} - \frac{Y_i\mathbb{I}(X_i=x)}{f(x|\mathbf{Z}_i)}\right|\right] \\
=& \mathbb{E}\left[\left|\frac{Y_1\mathbb{I}(X_1=x)}{\widehat{f}(x|\mathbf{Z}_1)} - \frac{Y_1\mathbb{I}(X_1=x)}{f(x|\mathbf{Z}_1)}\right|\right] && \text{Definição 3.20} \\
=& \mathbb{E}\left[\left|\frac{Y_i\mathbb{I}(X_i=x)(\widehat{f}(x|\mathbf{Z}_i)-f(x|\mathbf{Z}_i))}{\widehat{f}(x|\mathbf{Z}_i)f(x|\mathbf{Z}_i)}\right|\right] \\
\leq& \delta^{-2}\mathbb{E}\left[|Y_i\mathbb{I}(X_i=x)(\widehat{f}(x|\mathbf{Z}_i)-f(x|\mathbf{Z}_i))|\right] && \inf_z \min\{f(x|\mathbf{Z}_1),\widehat{f}(x|\mathbf{Z}_1)\} > \delta \\
=& \delta^{-2}\mathbb{E}\left[|\widehat{f}(x|\mathbf{Z}_i)-f(x|\mathbf{Z}_i)|\cdot\mathbb{E}[|Y_i\mathbb{I}(X_i=x)||\mathbf{Z}]\right] && \text{Lei da esperança total} \\
\leq& M\delta^{-2}\mathbb{E}\left[|\widehat{f}(x|\mathbf{Z}_i)-f(x|\mathbf{Z}_i)|\right] && \sup_z \mathbb{E}[|Y_i\mathbb{I}(X_i=x)|\mathbf{Z}=\mathbf{z}] < M \\
=& o(1)
\end{aligned}$$

□

A.2.5. Relativas ao Teorema 3.29

Prova do Teorema 3.29. Se as condições do Teorema 3.23 estão satisfeitas, então decorre deste resultado que $\widehat{\mathbb{E}}_1[Y|do(X=x)] \xrightarrow{\mathbb{P}} \mathbb{E}[Y|do(X=x)]$. Portanto, usando Lema A.12, resta demonstrar que

$$\mathbb{E}\left[\left|\widehat{\mathbb{E}}_2[Y|do(X=x)] - \sum_{i=1}^{n} \frac{\mathbb{I}(X_i=x)\hat{\mu}(x,\mathbf{Z}_i)}{n\hat{f}(x|\mathbf{Z}_i)}\right|\right] = o(1)$$

$$\begin{aligned}
&\mathbb{E}\left[\left|\widehat{\mathbb{E}}_2[Y|do(X=x)] - \sum_{i=1}^{n} \frac{\mathbb{I}(X_i=x)\hat{\mu}(x,\mathbf{Z}_i)}{n\hat{f}(x|\mathbf{Z}_i)}\right|\right] \\
=&\mathbb{E}\left[\left|\sum_{i=1}^{n} \frac{\mathbb{I}(X_i=x)(Y_i-\hat{\mu}(x,\mathbf{Z}_i))}{n\hat{f}(x|\mathbf{Z}_i)}\right|\right] && \text{Definição 3.25} \\
\leq& n{-}1\sum_{i=1}^{n}\mathbb{E}\left[\left|\frac{\mathbb{I}(X_i=x)(Y_i-\hat{\mu}(x,\mathbf{Z}_i))}{\hat{f}(x|\mathbf{Z}_i)}\right|\right] \\
=&\mathbb{E}\left[\left|\frac{\mathbb{I}(X_1=x)(Y_1-\hat{\mu}(x,\mathbf{Z}_1))}{\hat{f}(x|\mathbf{Z}_1)}\right|\right] && \text{Definição 3.20} \\
\leq&\delta^{-1}\mathbb{E}\left[\left|\mathbb{I}(X_1=x)(Y_1-\hat{\mu}(x,\mathbf{Z}_1))\right|\right] && \inf_{\mathbf{z}} \widehat{f}(x|\mathbf{z}) > \delta \\
\leq&\delta^{-1}\mathbb{E}\left[\left|\mathbb{I}(X_1=x)(\mathbb{E}[Y_1|X_1,\mathbf{Z}_1]-\hat{\mu}(x,\mathbf{Z}_1))\right|\right] && \text{Lei da esperança total} \\
=&\delta^{-1}\mathbb{E}\left[\left|\mathbb{I}(X_1=x)(\mathbb{E}[Y_1|X_1=x,\mathbf{Z}_1]-\hat{\mu}(x,\mathbf{Z}_1))\right|\right] && \mathbb{I}(X_1=x)\mathbb{E}[Y_1|X_1,\mathbf{Z}_1] \equiv \mathbb{I}(X_1=x)\mathbb{E}[Y_1|X_1=x,\mathbf{Z}_1] \\
\leq&\delta^{-1}\mathbb{E}\left[\left|\mathbb{E}[Y_1|X_1=x,\mathbf{Z}_1]-\hat{\mu}(x,\mathbf{Z}_1)\right|\right] \\
=&\mathbb{E}\left[\left|\mu(x,\mathbf{Z}_1)-\hat{\mu}(x,\mathbf{Z}_1)\right|\right] = o(1)
\end{aligned}$$

A seguir, se as condições do Teorema 3.26 estão satisfeitas, então decorre deste resultado que $\widehat{\mathbb{E}}_2[Y|do(X=x)] \xrightarrow{\mathbb{P}} \mathbb{E}[Y|do(X=x)]$. Portanto, usando Lema A.12, resta demonstrar que

$$\mathbb{E}\left[\left|\widehat{\mathbb{E}}_1[Y|do(X=x)] - \sum_{i=1}^{n} \frac{\mathbb{I}(X_i=x)\hat{\mu}(x,\mathbf{Z}_i)}{n\hat{f}(x|\mathbf{Z}_i)}\right|\right] = o(1)$$

$$
\begin{aligned}
&\mathbb{E}\left[\left|\widehat{\mathbb{E}}_1[Y|do(X=x)] - \sum_{i=1}^{n}\frac{\mathbb{I}(X_i=x)\hat{\mu}(x,\mathbf{Z}_i)}{n\hat{f}(x|\mathbf{Z}_i)}\right|\right] \\
=&\mathbb{E}\left[\left|\sum_{i=1}^{n}\frac{(\widehat{f}(x|\mathbf{Z}_i)-\mathbb{I}(X_i=x))\hat{\mu}(x,\mathbf{Z}_i)}{n\widehat{f}(x|\mathbf{Z}_i)}\right|\right] && \text{Definição 3.22} \\
\leq& n^{-1}\sum_{i=1}^{n}\mathbb{E}\left[\left|\frac{(\widehat{f}(x|\mathbf{Z}_i)-\mathbb{I}(X_i=x))\hat{\mu}(x,\mathbf{Z}_i)}{\widehat{f}(x|\mathbf{Z}_i)}\right|\right] \\
=&\mathbb{E}\left[\left|\frac{(\widehat{f}(x|\mathbf{Z}_1)-\mathbb{I}(X_1=x))\hat{\mu}(x,\mathbf{Z}_1)}{\widehat{f}(x|\mathbf{Z}_1)}\right|\right] && \text{Definição 3.20} \\
\leq&\delta^{-1}\mathbb{E}\left[\left|(\widehat{f}(x|\mathbf{Z}_1)-\mathbb{I}(X_1=x))\hat{\mu}(x,\mathbf{Z}_1)\right|\right] && \inf_{\mathbf{z}}\widehat{f}(x|\mathbf{Z}_1)>\delta \\
\leq&\delta^{-1}M\mathbb{E}\left[\left|\widehat{f}(x|\mathbf{Z}_1)-\mathbb{I}(X_1=x)\right|\right] && \sup_{\mathbf{z}}\hat{\mu}(x,\mathbf{z})<M \\
=&\delta^{-1}M\mathbb{E}\left[\left|\widehat{f}(x|\mathbf{Z}_1)-\mathbb{E}[\mathbb{I}(X_1=x)|\mathbf{Z}_1]\right|\right] && \text{Lei da esperança total} \\
=&\delta^{-1}M\mathbb{E}\left[\left|\widehat{f}(x|\mathbf{Z}_1)-f(x|\mathbf{Z}_1)\right|\right]=o(1)
\end{aligned}
$$

□

A.2.6. Relativas ao Teorema 3.35

Prova do Teorema 3.35. Se $X \equiv \mathbb{I}(\mathbf{Z} \geq \mathbf{z}_1)$, então:

$$
\begin{aligned}
CACE(\mathbf{Z}=\mathbf{z}_1) &= \mathbb{E}[Y|do(X=1),\mathbf{Z}=\mathbf{z}_1] - \mathbb{E}[Y|do(X=0),\mathbf{Z}=\mathbf{z}_1] && \text{Definição 3.5} \\
&= \lim_{\mathbf{z}\downarrow\mathbf{z}_1}\mathbb{E}[Y|do(X=1),\mathbf{Z}=\mathbf{z}] - \lim_{\mathbf{z}\uparrow\mathbf{z}_1}\mathbb{E}[Y|do(X=0),\mathbf{Z}=\mathbf{z}] && \text{continuidade} \\
&= \lim_{\mathbf{z}\downarrow\mathbf{z}_1}\mathbb{E}[Y|X=1,\mathbf{Z}=\mathbf{z}] - \lim_{\mathbf{z}\uparrow\mathbf{z}_1}\mathbb{E}[Y|X=0,\mathbf{Z}=\mathbf{z}] && \text{Teorema 3.18} \\
&= \lim_{\mathbf{z}\downarrow\mathbf{z}_1}\mathbb{E}[Y|X=1,\mathbf{Z}=\mathbf{z}] - \lim_{\mathbf{z}\uparrow\mathbf{z}_1}\mathbb{E}[Y|X=0,\mathbf{Z}=\mathbf{z}] && X\equiv\mathbb{I}(\mathbf{Z}\geq\mathbf{z}_1)
\end{aligned}
$$

A seguir, considere que $f(x|\mathbf{Z}) \in (0,1)$ é contínua exceto em $\mathbf{z}_1$. Primeiramente, note que

$$
\begin{aligned}
\mathbb{E}[Y|\mathbf{Z}] &= \mathbb{E}[\mathbb{E}[Y|X,\mathbf{Z}]|\mathbf{Z}] \\
&= \mathbb{E}[Y|X=1,\mathbf{Z}]f(X=1|\mathbf{Z}) + \mathbb{E}[Y|X=0,\mathbf{Z}](1-f(X=1|\mathbf{Z})) \\
&= (\mathbb{E}[Y|X=1,\mathbf{Z}] - \mathbb{E}[Y|X=0,\mathbf{Z}])f(X=1|\mathbf{Z}) + \mathbb{E}[Y|X=0,\mathbf{Z}] \\
&= CACE(\mathbf{Z})f(X=1|\mathbf{Z}) + \mathbb{E}[Y|do(X=0),\mathbf{Z}] && \text{Teorema 3.18} \qquad \text{(A.6)}
\end{aligned}
$$

Como $\mathbb{E}[Y|do(X=0),\mathbf{Z}]$ e $\mathbb{E}[Y|do(X=1),\mathbf{Z}]$ são contínuas em $\mathbf{z}_1$, $CACE(\mathbf{Z})$ também é contínua em $\mathbf{z}_1$. Assim, decorre da eq. (A.6) que

$$
\begin{aligned}
\lim_{\mathbf{z}\downarrow\mathbf{z}_1}\mathbb{E}[Y|\mathbf{Z}=\mathbf{z}] &= CACE(\mathbf{z}_1)\lim_{\mathbf{z}\downarrow\mathbf{z}_1}f(X=1|\mathbf{z}) + \mathbb{E}[Y|do(X=0),\mathbf{Z}=\mathbf{z}_1] \\
\lim_{\mathbf{z}\uparrow\mathbf{z}_1}\mathbb{E}[Y|\mathbf{Z}=\mathbf{z}] &= CACE(\mathbf{z}_1)\lim_{\mathbf{z}\uparrow\mathbf{z}_1}f(X=1|\mathbf{z}) + \mathbb{E}[Y|do(X=0),\mathbf{Z}=\mathbf{z}_1]
\end{aligned}
$$

Finalmente subtraindo as equações acima, obtemos

$$\lim_{\mathbf{z}\downarrow\mathbf{z}_1}\mathbb{E}[Y|\mathbf{Z}=\mathbf{z}] - \lim_{\mathbf{z}\uparrow\mathbf{z}_1}\mathbb{E}[Y|\mathbf{Z}=\mathbf{z}] = CACE(\mathbf{z}_1)(\lim_{\mathbf{z}\downarrow\mathbf{z}_1} f(X=1|\mathbf{z}) - \lim_{\mathbf{z}\uparrow\mathbf{z}_1} f(X=1|\mathbf{z}))$$
$$CACE(\mathbf{z}_1) = \frac{\lim_{\mathbf{z}\downarrow\mathbf{z}_1}\mathbb{E}[Y|\mathbf{Z}=\mathbf{z}] - \lim_{\mathbf{z}\uparrow\mathbf{z}_1}\mathbb{E}[Y|\mathbf{Z}=\mathbf{z}]}{\lim_{\mathbf{z}\downarrow\mathbf{z}_1} f(X=1|\mathbf{z}) - \lim_{\mathbf{z}\uparrow\mathbf{z}_1} f(X=1|\mathbf{z})}$$

□

A.3. Relativas às Seções 3.3 e 3.4

A.3.1. Relativas ao Teorema 3.44

Lema A.13. *$f^* \equiv f(\mathcal{V}|do(\mathbf{X}=\mathbf{x}))$ é compatível com $\mathcal{G}(\bar{\mathbf{X}})$. Além disso, $\mathbf{X}$ é degenerada em $\mathbf{x}$ segundo f^*.*

Demonstração. Decorre da Definição 3.1 que $f^*(\mathcal{V}) = \mathbb{I}(\mathbf{X}=\mathbf{x})\prod_{V\notin\mathbf{X}} f(V|Pa(V))$. Definindo $g_{X_i}(X_i) = \mathbb{I}(X_i = x_i)$, para todo $X_i \in \mathbf{X}$ e $g_V(V, Pa(V)) = f(v|Pa(V))$, note que

$$f^*(\mathcal{V}) = \prod_{X_i\in\mathbf{X}} g_{X_i}(X_i) \prod_{V\notin\mathbf{X}} g_V(V, Pa(V))$$

Portanto, decorre do Lema 2.11 que f^* é compatível com um grafo em que todo $X_i \in \mathbf{X}$ não tem pais e todo $V \notin \mathbf{X}$ tem os mesmos pais que em $\mathcal{G}$. Isto é, $\mathcal{G}$ é compatível com $\mathcal{G}(\bar{\mathbf{X}})$.

Além disso, tomando $\mathbb{V} = \mathcal{V} - \mathbf{X}$,

$$\begin{aligned}
f^*(\mathbf{X}) &= \int f^*(sV)d\mathbb{V} \\
&= \int \mathbb{I}(\mathbf{X}=\mathbf{x}) \prod_{V\in\mathbb{V}} f(V|Pa(V))d\mathbb{V} \\
&= \mathbb{I}(\mathbf{X}=\mathbf{x}) \int \prod_{V\in\mathbb{V}} f(V|Pa(V))d\mathbb{V} \\
&= \mathbb{I}(\mathbf{X}=\mathbf{x}).
\end{aligned}$$

Portanto, $\mathbf{X}$ é degenerado em $\mathbf{x}$ segundo f^*. □

Lema A.14. *Se $\mathbf{Y} \perp^d \mathbf{Z}|\mathbf{X}\cup\mathbf{W}$ em $\mathcal{G}(\bar{\mathbf{X}})$, então*

$$f(\mathbf{Y}|do(\mathbf{X}),\mathbf{Z},\mathbf{W}) = f(\mathbf{Y}|do(\mathbf{X}),\mathbf{W})$$

Demonstração. Seja $f^*(\mathcal{V}) \equiv f(\mathcal{V}|do(\mathbf{X}=\mathbf{x}))$. Decorre do Lema A.13 que f^* é compatível com $\mathcal{G}(\bar{\mathbf{X}})$ e que $\mathbf{X}$ é degenerado em $\mathbf{x}$ segundo f^*. Assim,

$$\begin{aligned}
f(\mathbf{Y}|do(\mathbf{X}),\mathbf{Z},\mathbf{W}) &= f^*(\mathbf{Y}|\mathbf{Z},\mathbf{W}) \\
&= f^*(\mathbf{Y}|\mathbf{X}=\mathbf{x},\mathbf{Z},\mathbf{W}) && \mathbf{X} \text{ é degenerado segundo } f^* \\
&= f^*(\mathbf{Y}|\mathbf{X}=\mathbf{x},\mathbf{W}) && f^* \text{ compatível com } \mathcal{G}(do(\mathbf{X})), \\
& && \mathbf{Y} \perp^d \mathbf{Z}|\mathbf{X}\cup\mathbf{W} \text{ em } \mathcal{G}(do(\mathbf{X})), \text{ e Teorema 2.40} \\
&= f(\mathbf{Y}|do(\mathbf{X}=\mathbf{x}),\mathbf{W}).
\end{aligned}$$

$\square$

Lema A.15. *Se* $Y \perp^d \mathbf{W}|\mathbf{Z} \cup \mathbf{X}$ *em* $\mathcal{G}(\bar{\mathbf{X}}, \underline{\mathbf{W}})$*, então*

$$f(Y|do(\mathbf{X}), do(\mathbf{W}), \mathbf{Z}) = f(Y|do(\mathbf{X}), \mathbf{W}, \mathbf{Z})$$

Demonstração. Seja $f^*(\mathcal{V}) \equiv f(\mathcal{V}|do(\mathbf{X} = \mathbf{x}))$. Como $Y \perp^d \mathbf{W}|\mathbf{Z} \cup \mathbf{X}$ em $\mathcal{G}(\bar{\mathbf{X}}, \underline{\mathbf{W}})$, não há nenhum caminho ativo em $\mathcal{G}(\bar{\mathbf{X}})$ de $\mathbf{X}$ em Y que inicia com $\mathbf{W} \leftarrow$. Isto é, $\mathbf{X} \cup \mathbf{Z}$ satisfaz o segundo item do critério backdoor para medir o efeito causal de $\mathbf{W}$ em Y. Portanto,

$$\begin{aligned}
f(Y|do(\mathbf{X} = \mathbf{x}), do(\mathbf{W}), \mathbf{Z}) &= f^*(Y|do(\mathbf{W}), \mathbf{Z}) \\
&= f^*(Y|do(\mathbf{W}), \mathbf{X} = \mathbf{x}, \mathbf{Z}) && \text{Lema A.13} \\
&= f^*(Y|\mathbf{W}, \mathbf{X} = \mathbf{x}, \mathbf{Z}) && \text{Lema A.8} \\
&= f^*(Y|\mathbf{W}, \mathbf{Z}) \\
&= f(Y|do(\mathbf{X} = \mathbf{x}), \mathbf{W}, \mathbf{Z})
\end{aligned}$$

$\square$

Lema A.16. *Se* $\mathbf{Y} \perp^d I_{\mathbf{X}}|\mathbf{Z} \cup \mathbf{W}$ *em* $\mathcal{G}(\bar{\mathbf{W}}, \mathbf{X}^+)$*, então:*

$$f(Y|do(\mathbf{W}), do(\mathbf{X}), \mathbf{Z}) = f(Y|do(\mathbf{W}), \mathbf{Z})$$

Demonstração. Seja $f^*(\mathcal{V}) \equiv f(\mathcal{V}|do(\mathbf{W} = \mathbf{w}))$.

$$\begin{aligned}
f(Y|do(\mathbf{W} = \mathbf{w}), do(\mathbf{X}), \mathbf{Z}) &= f^*(Y|do(\mathbf{X}), \mathbf{Z}) && \text{Lema A.13} \\
&= f^*(Y|do(\mathbf{X}), \mathbf{W} = \mathbf{w}, \mathbf{Z}) && \text{Lema A.13} \\
&= f^*_*(Y|I_{\mathbf{X}} = 1, \mathbf{W} = \mathbf{w}, \mathbf{Z}) && \text{Lema A.5} \\
&= f^*_*(Y|\mathbf{W} = \mathbf{w}, \mathbf{Z}, I_{\mathbf{X}} = 0) && Y \perp^d I_{\mathbf{X}}|\mathbf{W} \cup \mathbf{Z} \text{ em } \mathcal{G}(\bar{\mathbf{W}}, \mathbf{X}^+) \\
&= f^*(Y|\mathbf{W} = \mathbf{w}, \mathbf{Z}) && \text{Lema A.6} \\
&= f(Y|do(\mathbf{W} = \mathbf{w}), \mathbf{Z}) && \text{Lema A.13}
\end{aligned}$$

$\square$

Prova do Teorema 3.44. Decorre dos Lemas A.14 a A.16. $\square$

A.3.2. Relativas ao Teorema 3.40

Lema A.17. *Se* $\mathbf{W}$ *satisfaz o critério frontdoor para medir o efeito causal de* X *em* Y*, então* $f(Y|do(X), \mathbf{W}) = f(Y|do(X), do(\mathbf{W}))$.

Demonstração. Decorre do critério frontdoor Definição 3.39.3 que X satisfaz o item 2 do critério backdoor para medir o efeito causal de $\mathbf{W}$ em Y. Portanto, pelo Lema 3.45, $\mathbf{Y} \perp^d \mathbf{W}|\mathbf{X}$ em $\mathcal{G}(\underline{\mathbf{W}})$. Pelo Exercício 2.47, $\mathbf{Y} \perp^d \mathbf{W}|\mathbf{X}$ em $\mathcal{G}(\bar{\mathbf{X}}, \underline{\mathbf{W}})$. A prova se conclui aplicando o item 2 do Teorema 3.44. $\square$

Lema A.18. *Se* $\mathbf{W}$ *satisfaz o critério frontdoor para medir o efeito causal de* X *em* Y*, então* $f(Y|do(X), do(\mathbf{W})) = f(Y|do(\mathbf{W}))$.

Demonstração. A prova consiste em aplicar o item 3 do Teorema 3.44. Para tal, desejamos provar que $\mathbf{Y} \perp^d I_X | \mathbf{W}$ em $\mathcal{G}(\bar{\mathbf{W}}, X^+)$. Tome C como um caminho arbitrário em $\mathcal{G}(\bar{\mathbf{W}}, X^+)$ de $I_{\mathbf{X}}$ em Y.

Primeiramente, provaremos que C não é um caminho direcionado. C não é um caminho direcionado de Y a I_X pois a única aresta em $\mathcal{G}(\bar{\mathbf{W}}, X^+)$ ligada a I_X é $I_X \rightarrow X$. A seguir, suponha que C é um caminho direcionado de I_X a Y. Pelo Definição 3.39.2, existe C_i tal que $C_i \in \mathbf{W}$. Como C é direcionado de I_X em Y, $C_{i-1} \rightarrow C_i$. Este é um absurdo, pois não há aresta apontando para $\mathbf{W}$ em $\mathcal{G}(\bar{\mathbf{W}}, X^+)$. Portanto, C não é um caminho direcionado.

Conclua que existe C_i que é um colisor. Como não há arestas apontando para $\mathbf{W}$ em $\mathcal{G}(\bar{\mathbf{W}}, X^+)$, $C_i \notin \mathbf{W}$. Como C_i é um colisor e $C_i \notin \mathbf{W}$, conclua que C está bloqueado. Como C era arbitrário, $\mathbf{Y} \perp^d I_X | \mathbf{W}$ em $\mathcal{G}(\bar{\mathbf{W}}, X^+)$. □

Prova do Teorema 3.40.

$$
\begin{aligned}
f(Y|do(X=x)) &= \int f(Y|do(X=x), \mathbf{W}) f(\mathbf{W}|do(X=x)) d\mathbf{W} \\
&= \int f(Y|do(X=x), \mathbf{W}) f(\mathbf{W}|X=x) d\mathbf{W} && \text{Exercício 3.48} \\
&= \int f(Y|do(X=x), do(\mathbf{W})) f(\mathbf{W}|X=x) d\mathbf{W} && \text{Lema A.17} \\
&= \int f(Y|do(\mathbf{W})) f(\mathbf{W}|X=x) d\mathbf{W} && \text{Lema A.18} \\
&= \int \int f(Y|\mathbf{W}, X) f(X) dX f(\mathbf{W}|X=x) d\mathbf{W} && \text{Exercício 3.48}
\end{aligned}
$$

□

A.3.3. Relativas ao Teorema 3.41

Prova do Teorema 3.41.

$$
\begin{aligned}
\mathbb{E}[Y|do(X=x)] &= \int Y f(Y|do(X=x)) dY \\
&= \int Y \int f(\mathbf{W}|x) \int f(Y|X, \mathbf{W}) f(X) dX dW dY && \text{Teorema 3.40} \\
&= \int Y f(Y|X, \mathbf{W}) f(X) f(\mathbf{W}|x) d(Y \times \mathbf{W} \times Y) \\
&= \int \frac{Y f(\mathbf{W}|x)}{f(\mathbf{W}|X)} f(Y, \mathbf{W}, X) d(Y \times \mathbf{W} \times Y) \\
&= \mathbb{E}\left[\frac{Y f(\mathbf{W}|x)}{f(\mathbf{W}|X)}\right]
\end{aligned}
$$

□