Instituto de MATEMÁTICA E ESTATÍSTICA UFRGS

DEPARTAMENTO DE ESTATÍSTICA
E-mail: dest@mat.ufrgs.br

Trabalho de Conclusão de Curso

Avaliação de Consistência DAG-dataset: aplicação em um estudo epidemiológico

Bruna Silveira da Rosa

2 de junho de 2021

Bruna Silveira da Rosa

Avaliação de Consistência DAG-dataset: aplicação em um estudo epidemiológico

Trabalho de Conclusão apresentado à comissão de Graduação do Departamento de Estatística da Universidade Federal do Rio Grande do Sul, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

Orientador: Prof. Dr. Rodrigo Citton Padilha dos Reis

Porto Alegre
Maio de 2021

Bruna Silveira da Rosa

Avaliação de Consistência DAG-dataset: aplicação em um estudo epidemiológico

Este Trabalho foi julgado adequado para obtenção dos créditos da disciplina Trabalho de Conclusão de Curso em Estatística e aprovado em sua forma final pela Orientador e pela Banca Examinadora.

Orientador:_____
Prof. Dr. Rodrigo Citton Padilha dos Reis, UFRGS
Doutor pela Universidade Federal de Minas Gerais, Belo Horizonte, MG

Banca Examinadora:

Prof[a]. Dr[a]. Márcia Helena Barbian, UFRGS
Doutora pela Universidade Federal de Minas Gerais, Belo Horizonte, MG

Porto Alegre
Maio de 2021

"Os que se encantam com a prática sem a ciência são como os timoneiros que entram no navio sem timão nem bússola, nunca tendo certeza do seu destino".
(Leonardo da Vinci)

Agradecimentos

Aos meus pais, Susana e Pedro, por todo amor, incentivo e por sempre acreditarem em mim.

À minha irmã Daniela, por me ajudar muito nessa reta final de curso, sempre estando disponível.

À todos meus familiares, pelo apoio e carinho de sempre.

À Larissa, por sempre estar perto, mesmo que longe.

Aos meus amigos do tempo de escola, Bruna do Prado, Kimberly, Guilherme, Melany, Isis, Camila e Bárbara, ter vocês comigo por todos esses anos é um privilégio. É como diz aquele famoso ditado: O que o Gentil uniu, ninguém separa!

À Gabriela e Juliana, por estarem comigo desde o primeiro ano de curso, ter vocês do meu lado ao longo desses anos foi fundamental para chegar até aqui. À Gabriela, agradeço especialmente por sempre compreender as minhas inseguranças e me dar forças para continuar. À Juliana, por ser a melhor professora particular desse mundo. Obrigada por tudo!

Ao Renan, que compartilhou comigo tantas horas de espera pelo TM1, que tantas vezes fez comigo o caminho da Bento até as aulas. Obrigada por fazer esses momentos valerem a pena de serem lembrados.

À Franciele e Giulia, por tantas vezes serem minhas parceiras nos trabalhos, e sempre estarem dispostas para tudo. Sou muito feliz de ter a amizade de vocês.

À Maitê, por ser um exemplo de foco e perseverança. Tenho muita admiração por você.

Aos demais colegas da Estatística, por toda parceria e cumplicidade.

Ao meu orientador, professor Rodrigo, por todo ensinamento, paciência e gentileza durante este último ano.

À todos os professores do Departamento de Estatística que contribuíram para minha formação.

Por fim, agradeço a UFRGS, por me proporcionar essa experiência, e realizar um sonho. Pelo ensino de qualidade, e por todas as pessoas que colocou no meu caminho durante esses anos.

Resumo

Este trabalho apresenta e demonstra como é realizada a avaliação de consistência *DAG-dataset*, visto que nem todo DAG é consistente com o conjunto de dados gerado pelo processo causal. Primeiramente, serão apresentados conceitos importantes para o entendimento da avaliação de consistência *DAG-datase*, como a terminologia e construção dos DAGs, Independência Condicional, Relações de Independência, Implicações Testáveis e Consistência *DAG-daset*. Então, uma breve introdução ao uso das ferramentas *'DAGitty'* Web e ao pacote `dagitty`, será feita. Essas ferramentas auxíliam na hora em que se está trabalhando com diagramas causais, tanto para desenhar o DAG, quanto para analisar as relações causais deste, especialmente quando se está trabalhando com DAGs grandes, de alta complexidade. Sendo assim, possível realizar a avaliação de consistência *DAG-dataset* por meio de uma das ferramentas. O funcionamento dessas ferramentas será apresentado através da aplicação de um estudo epidemiológico, onde o interesse é avaliar a associação putativa do nível de catecolaminas endógenas no sangue com a incidência decorrente de doença cardíaca coronária.

Palavras-Chave: Avaliação de Consistência, DAGs, Diagramas Causais, d-separação, Grafos Acíclicos Dirigidos.

Abstract

This work presents and demonstrates how the evaluating *DAG-dataset* consistency is performed, since not all DAG is consistent with the data set generated by the causal process. First, important concepts will be presented for understanding the consistency evaluating of the *DAG-datase*, such as the terminology and construction of the DAGs, Conditional Independence, Independence Relationships, Testable Implications and Consistency *DAG-daset*. Then, a brief introduction to the use of the *'DAGitty'* Web and the `dagitty` package tools will be made. These tools help when working with causal diagrams, both to design the DAG and to analyze its causal relationships, especially when working with large, highly complex DAGs. Therefore, it is possible to carry out a evaluating *DAG-dataset* consistency using one of the tools. The functioning of these tools will be presented through the application of an epidemiological study, where the interest is to evaluate the putative association of the level of endogenous catecholamines in the blood with the incidence resulting from coronary heart disease.

Keywords: Causal Diagrams, DAGs, Directed Acyclic Graphs, d-separation, Evaluating Consistency.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

Os Grafos Acíclicos Dirigidos (DAGs, sigla em inglês para *directed acyclic graphs*) são ferramentas gráficas comumente utilizadas para estabelecer relações causais entre variáveis. Podem ser aplicados em diversas áreas do conhecimento ou de pesquisa. Como nas análises de inferência causal em epidemiologia, onde desenhar um DAG é uma das principais estratégias usadas para expressar as hipóteses sobre as relações entre variáveis e minimizar o viés de confusão (Textor et al., 2016). Os DAGs também são chamados de diagrama de influência, diagrama de relevência ou rede causal, são modelos gráficos que podem desempenhar um papel complementar aos modelos estatísticos convencionais, que incorporam muitas suposições paramétricas que podem ser consideradas incorretas (Greenland et al., 1999). Ou seja, os DAGs não incorporam as fortes suposições paramétricas dos modelos convencionais, sendo assim, possível exibir as suposições sobre a rede de causalidade que não são obtidas por tais modelos (Greenland et al., 1999).

Desde o início do século XX, os modelos de sistemas causais, com o auxílio de representações gráficas ou diagramas de caminho, são utilizados (Wright, 1920). Mas somente a partir dos anos 1990, que os diagramas começaram a ser empregados como ferramenta no campo da inferência causal (Greenland e Pearl, 2007).

Estes modelos gráficos acabam revelando certas deficiências despercebidas pelos critérios tradicionais, quando usados na consideração de múltiplos fatores de confusão em potencial (Greenland et al., 1999), por esses motivos o uso de DAGs vem crescendo nas pesquisas epidemiológicas. Desta maneira, tornou-se uma ferramenta popular na hora de identificar variáveis de confusão que precisam ser ajustadas na estimativa de efeitos causais (Tennant et al., 2020). Uma forma de obter as estimativas dos efeitos causais é por meio de um experimento controlodado randomizado. Porém, na maior parte dos casos, os efeitos causais devem ser estimados a partir de dados observacionais, já que restrições práticas e éticas impossibilitam a realização de um estudo experimental controlado e randomizado (Tennant et al., 2020). Por exemplo, um estudo que aleatoriamente alocasse indivíduos em grupos de fumantes (provavelmente submetendo indivíduos não fumantes ao ato de fumar) e não

fumantes (provavelmente submento indivíduos fumantes à privação do ato de fumar), para estudar o efeito do tabagismo em agravos de saúde (por exemplo, câncer de pulmão), violaria princípios éticos. Assim, a relação causal entre tabagismo e agravos de saúde é geralmente estudada, e portanto, estimada a partir de estudos observacionais. Consequentemente, variadas fontes de viés surgem quando se está trabalhando com dados observacionais (Tennant et al., 2020).

Assim, abordagens de inferência causal são muito utilizadas, como por exemplo a Estrutura de Desfechos Potenciais, visto que proporcionam maior transparência quando se está lidando com dados observacionais para as estimativas de efeitos causais (Tennant et al., 2020). Desta forma, é possível definir os efeitos causais de interesse antes do início das análises. Isso é aceitável desde que se tenha o conhecimento externo do processo de gereção de dados e o controle sobre os fatores de confusão, ou seja, as causas mútuas entre exposição e desfecho (Tennant et al., 2020).

Uma questão essencial à ser avaliada é a consistência *DAG-dataset*, para verificar se os dados gerados pelo processo causal estão corretos, ou seja, verificar se o modelo causal estabelecido pelo DAG é consistente com os dados observados. A avalição de consistência *DAG-dataset* é feita através da análise de uma série de relações de independência condicional entre as variáveis do DAG que foram impostas pela estrutura do modelo causal. Há disponível um aplicativo na Web e um pacote para o software R chamado *'DAGitty'*, desenvolvido por Textor et al. (2016), para desenhar DAGs e avaliar a consistência *DAG-dataset*. Atualmente, no pacote `dagitty`, disponível para uso no software R, só é possível trabalhar com variáveis normalmente distribuídas, oferecendo unicamente testes paramétricos e semi-paramétricos de independência condicional, suportando o uso de regressão linear e regressão polinomial local para calcular os resíduos para o teste de independência condicional (Textor et al., 2016).

Apesar do crescente uso dos DAGs nas pesquisas epidemiológicas, ferramenta utilizada no auxílio da detecção de confundimento entre variáveis, viés de seleção e viés de informação, os DAGs ainda são uma ferramenta pouco utilizada. Um dos possíveis motivos é que certos temas abordados em programas de investigação levam em conta o grau de incerteza sobre os mecanismos dos processos de geração de dados (Cortes et al., 2016).

Tennant et al. (2020) apresentam informações muito interessantes sobre o uso dos DAGs na pesquisa em saúde. Em seu estudo de revisão foram analisados 234 artigos que fizeram uso desses modelos gráficos. Os dois países com maior número de publicações foram Estados Unidos 34% (n=79) e Alemanha 12% (n=29). Em 62% (n=144) dos artigos os DAGs foram apresentados, ou no próprio trabalho ou em algum link disponível. Nos outros 38% (n=90) dos artigos nenhum DAG foi exposto. Ainda, nenhum dos artigos relatou avaliar a compatibilidade de seus DAGs

com o conjunto de dados. Apenas 21% (n=48) desses artigos relatam as estimativas causais de interesse, 8% (n=18) estavam interessados nos efeitos causais totais, 8% (n=18) nos efeitos múltiplos e 5% (n=12) buscaram os efeitos causais diretos. A especificação desses estimadores auxiliam na clareza dos objetivos do estudo e nas interpretações do modelo, mas poucos artigos apresentam essas informações. É pontuado que isso muitas vezes ocorre pela falta de incentivo e falta de confiança por parte dos pesquisadores pelo fato dos DAGs serem uma ferramenta ainda pouco aplicada em pesquisas epidemiológicas.

O objetivo deste trabalho é introduzir a terminologia e notação dos DAGs, apresentar as propriedades de independência condicional induzidas pela configuração de um DAG e os testes de hipóteses para avaliação consistência DAG-dataset implementados no pacote `dagitty`.

Esta monografia está organizada da seguinte forma: o Capítulo 2 apresenta uma introdução sobre independência condicional, DAGs, como a probabilidade e os DAGs estão relacionados, relações de independência e os passos na construção de um DAG. No Capítulo 3 é abordado a consistência *DAG-dataset*: quais as implicações testáveis, a avaliação de consistência *DAG-dataset*, os testes de independência condicional que foram considerados, e classes equivalentes. O Capítulo 4 apresenta uma breve introdução ao *'DAGitty'* Web e aos pacotes `dagitty` e `ggdag`, a interface do aplicativo, suas principais funções, como desenhar um DAG no *DAGitty* Web, fazer a avaliação de consistência *DAG-dataset* com o pacote `dagitty`, alterar o *layout* dos grafos através do pacote `ggdag`; e a demonstração do uso dessas funções através da aplicação de um estudo epidemiológico, apresentado em Kleinbaum et al. (1982), onde é avaliada a associação putativa do nível de catecolaminas endógenas no sangue (exposição) com a incidência decorrente de doença cardíaca coronária (desfecho). Por fim, o Capítulo 5 apresenta as conclusões obtidas com este trabalho e sugestões de futuros tópicos de pesquisa que podem vir a ser desenvolvidos a partir desta monografia.

2 DAGs e Independência Condicional

Os conceitos de dependência, independência e independência condicional são fundamentais na área da estatística. Aqui, iremos nos ater aos conceitos de independência condicional, fundamentais para a avaliação de consistência *DAG-dataset*.

Segundo Textor et al. (2016), podemos fazer a avaliação de consistência *DAG-dataset* através de restrições estatisticamente testáveis que são consequência da propriedade de *d-separação*, que impõe restrições no conjunto de dados gerado pelo processo causal descrito pelo DAG, como independência condicional e incondicional.

As relações de independência condicional são um elemento importante para responder as questões de inferência causal, que tem tido uma grande crescente quando o interesse de pesquisadores é avaliar a relação de variáveis e como essas interagem (Li e Fan, 2019).

2.1 Independência Condicional

Por definição, duas variáveis aleatórias X e Y são condicionalmente independentes dada uma terceira variável aleatória $Z (X \perp Y \mid Z)$, se

$$f_{X,Y|Z}(x, y \mid z) = f_{X|Z}(x \mid z) f_{Y|Z}(y \mid z), \quad \forall x, y \; e \; z.$$

Uma definição equivalente a essa é

$$f(x \mid y, z) = f(x \mid z).$$

Ou seja, significa que uma vez que Z é conhecido, Y não nos fornece mais informações sobre X.

Algumas propriedades são válidas para as relações de independência condicional, como:

$$X \perp Y \mid Z \Longrightarrow Y \perp X \mid Z$$
$$X \perp Y \mid Z \; e \; U = h(X) \Longrightarrow U \perp Y \mid Z$$
$$X \perp Y \mid Z \; e \; U = h(X) \Longrightarrow X \perp Y \mid (Z, U)$$
$$X \perp Y \mid Z \; e \; X \perp W \mid (Y, Z) \Longrightarrow X \perp (W, Y) \mid Z$$
$$X \perp Y \mid Z \; e \; X \perp Z \mid Y \Longrightarrow X \perp (Y, Z).$$

Shipley (2002) comenta como a relação de causalidade e a independência condicional entre variáveis pode se tornar confusa muitas vezes. Isso ocorre porque, mesmo duas variáveis sendo dependentes e correlacionas elas podem ser condicionalmente independentes dado outro conjunto de variáveis. Essa relação se dá através da propriedade *d-separação*, que será discutida na Seção 2.5.

2.2 DAGs: notação e terminologia

Um Grafo Acíclico Dirigido pode ser construído abstraindo as suposições causais incluídas em uma descrição das relações hipotéticas entre as variáveis de estudo (Greenland et al., 1999). A terminologia dos DAGs apresentada nesta seção segue a terminologia adotada por Wasserman (2004) e Tennant et al. (2020).

Consideremos que um grafo acíclico dirigido G consiste em um conjunto de vértices (ou nós) V, que representam as variáveis e suas medidas, e em um conjunto de arestas (ou setas) A, que representam as relações hipotéticas entre os vértices. Logo, se duas variáveis $(X, Y) \in A$, haverá uma seta apontando de X para Y, como na Figura 2.1.

Figura 2.1: Um DAG com V = {X,Y} e A = {(X,Y)}.

As arestas são unidirecionais, portanto o grafo é dirigido. Um vértice não pode causar a si mesmo, porque não é possível que uma variável cause a si mesma em um dado momento no tempo, ou seja, o grafo é acíclico.

Uma aresta entre dois véstices indica que existe uma relação causal entre as variáveis. Se houver uma aresta de X para Y, então X é um **pai** (ou **mãe**) de Y e Y é um **filho** (ou **filha**) de X. O conjunto de pais de X pode ser denotado por π_X ou $\pi(X)$. Dizemos que um caminho é um conjunto de uma ou mais arestas que conectam dois vértices. Esse caminho pode ser aberto ou fechado. Um caminho aberto informa que existe associação estatística entre variáveis, enquanto o caminho fechado implica em ausência de associação entre variáveis.

Em um caminho dirigido, o conjunto de arestas que conectam dois vértices devem apontar na mesma direção. Então, se houver um caminho dirigido de X para Z como na Figura 2.2, podemos dizer que X é um ancestral de Z, e que Z é um descendente de X.

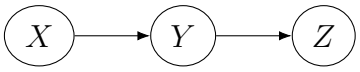

Figura 2.2: Caminho dirigido.

Também é possível dizer que X é um ancestral de Y, e que Y é um descendente de X. Ou seja, pais são ancestrais, e filhos são descendentes.

Um caminho com um colisor, é um caminho fechado (ou bloqueado) entre a exposição e o desfecho que passa por um ou mais colisores. Um colisor é um vértice que recebe duas ou mais arestas, por exemplo, X é a exposição e Z é o desfecho, ambos causam Y (Figura 2.3).

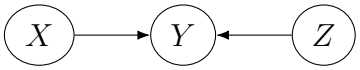

Figura 2.3: Caminho com colisor.

2.3 Construção de DAGs

Um caminho com um ou mais colisores pode introduzir certo viés de colisão. Portanto, quando quisermos determinar se a inclusão de uma covariável no conjunto de ajuste irá aumentar ou diminuir o viés na estimativa de efeito, é importante termos conhecimento da estrutura causal (Cortes et al., 2016). Podemos fazer essa verificação através do conjunto de ajustes suficiente S, que fechará todos os caminhos com viés de confusão, deixando somente os caminhos causais abertos. Ou seja, para controlar o confundimento é preciso selecionar um conjunto de covariáveis, denominado por S, que fechará todos os caminhos pelo critério *back-door* e manterá todos os caminhos causais entre a exposição e desfecho (Cortes et al., 2016).

Será utilizado o exemplo apresentado em Cortes et al. (2016) de um estudo epidemiológico, em que a variável `estresse no trabalho` é considerada a exposição e a variável `obesidade` o desfecho, para verificarmos como é possível encontrar o conjunto de ajustes suficiente S para o DAG do estudo. Mas antes, vamos desenhar o DAG que corresponde ao exemplo, seguindo os seguintes passos.

1. Escrever a exposição e o desfecho, ligando esses por uma seta na Figura 2.4.

   A aresta representa o efeito causal que queremos estimar. No caso deste exemplo, o efeito total do `estresse no trabalho` na `obesidade`.

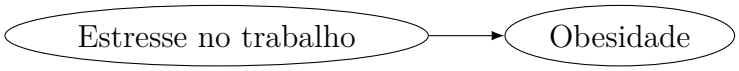

Figura 2.4: Construção do DAG: relação da exposição e desfecho.

2. Acrescentamos as causas comuns entre a exposição e o desfecho na Figura 2.5.

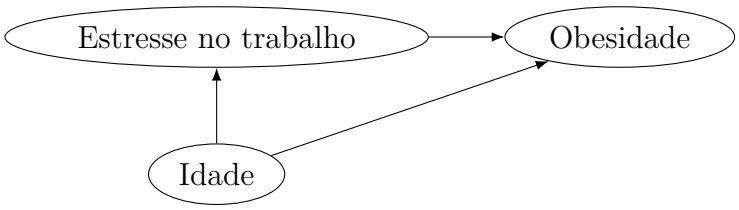

Figura 2.5: Construção do DAG: causa comum.

3. Então, continuamos acrescentando qualquer variável que seja uma causa comum entre duas ou mais variáveis.

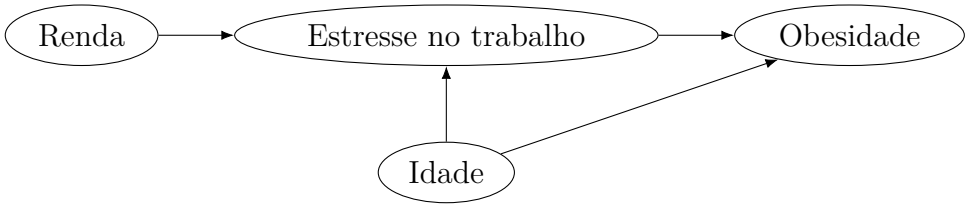

Figura 2.6: Construção do DAG: adição da variável `Renda`.

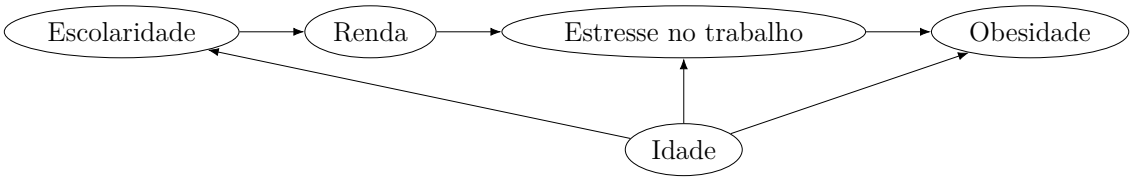

Figura 2.7: Construção do DAG: adição da variável `Escolaridade`.

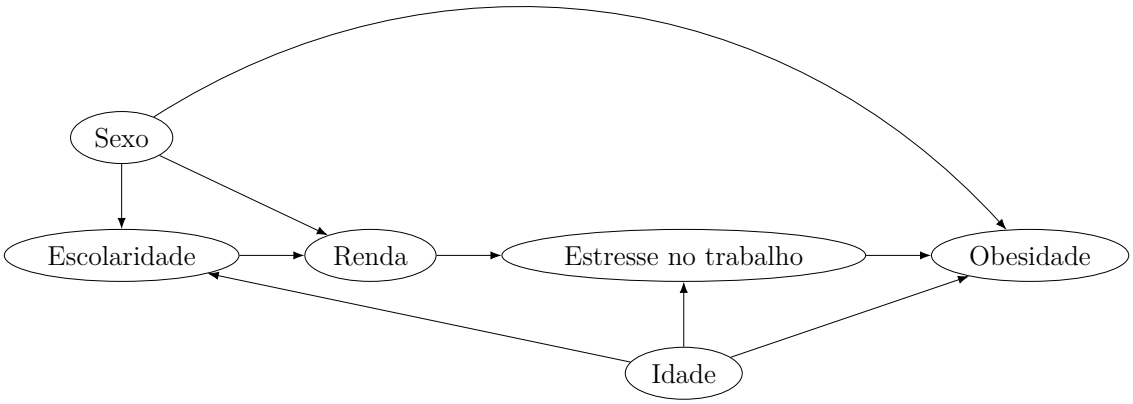

Figura 2.8: Construção do DAG: adição da variável `Sexo`.

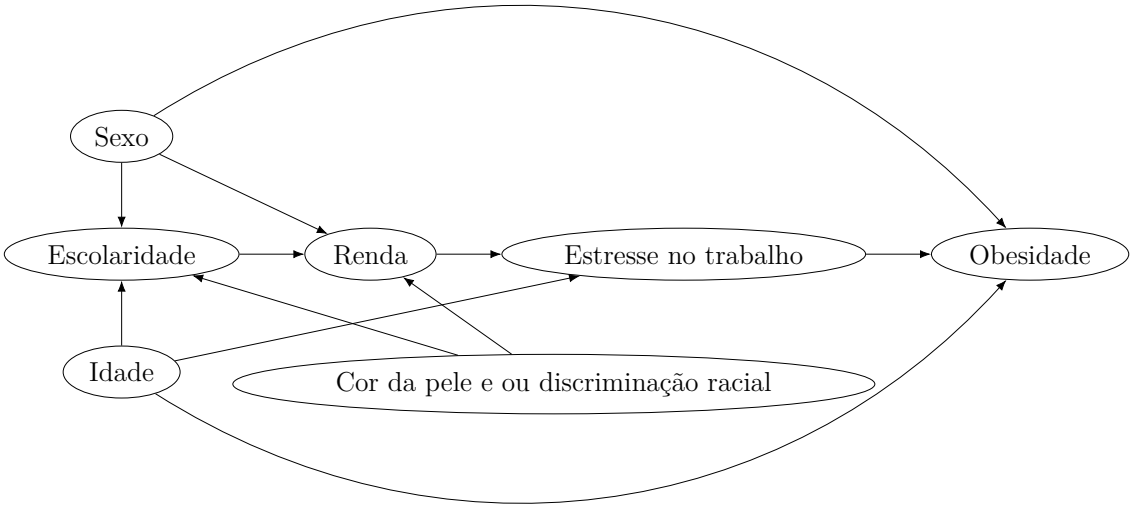

Figura 2.9: Construção do DAG: adição da variável `Cor da pele e ou discriminação racial`.

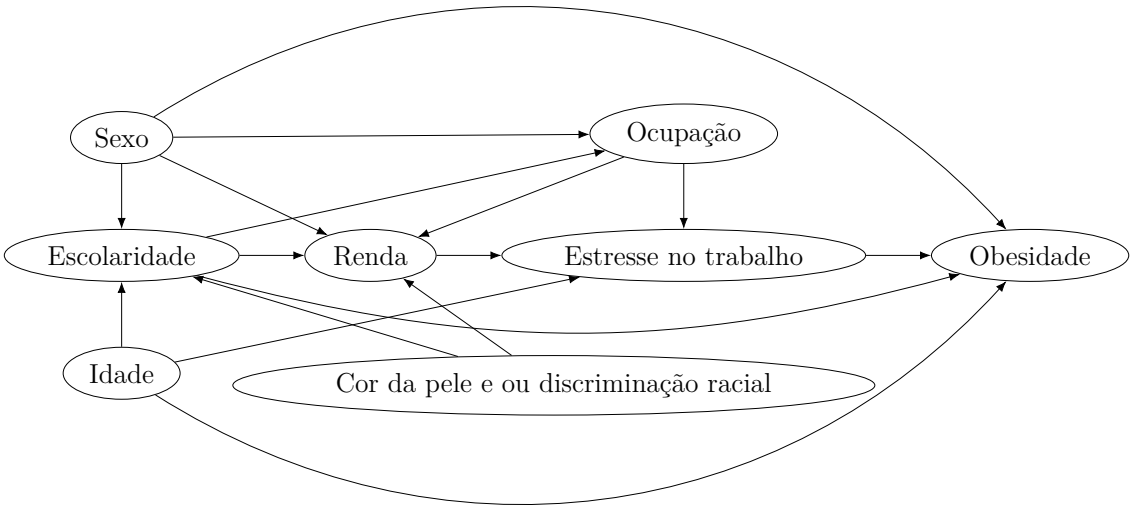

Figura 2.10: Construção do DAG: adição da variável `Ocupação`. Exemplo de um DAG representando as relações de causa e efeito entre variáveis. Exemplo apresentado em Cortes et al. (2016).

As Figuras 2.4, 2.5, 2.6, 2.7, 2.8, 2.9 e 2.10 mostram como cada variável é adicionada ao grafo, até por fim termos o DAG final, que representa a estrutura causal teórica formulada pelo pesquisador.

Greenland et al. (1999) apresentam um algoritmo para se chegar ao conjunto de ajustes suficiente S, por meio do critério *back-door*:

(i) Remova todas as arestas originadas da exposição. Assim, todos os efeitos da exposição serão removidos, visto que o conjunto S não deve conter nenhum descendente da exposição, pois quando condicionado pelos efeitos da exposição, é possível desbloquear caminhos da exposição para o desfecho que podem introduzir certo viés.

(ii) Em seguida, com uma linha tracejada, junte todos os vértices que compartilham um filho presente no conjunto S, ou que tenha um descendente no conjunto S. A linha tracejada induz novas associações entre as variáveis dado o conjunto de ajustes suficiente S.

(iii) Por último, verifique no novo gráfico se existe algum caminho aberto da exposição para o desfecho que não passe por nenhum elemento do conjunto S. Então, se todos os caminhos abertos forem interceptados por uma variável presente no conjunto S, o conjunto S será suficiente para controlar o confundimento.

Dado o grafo acíclico dirigido (Figura 2.10), podemos avaliar, através do critério *back-door*, se as variáveis sexo, renda e idade são suficientes para controlar o confundimento.

Por (i) e (ii) remova os efeitos da variável de exposição e faça uma linha tracejada nas variáveis presentes no conjunto S que possuam descendetes em comum ou um filho no conjunto (Figura 2.11).

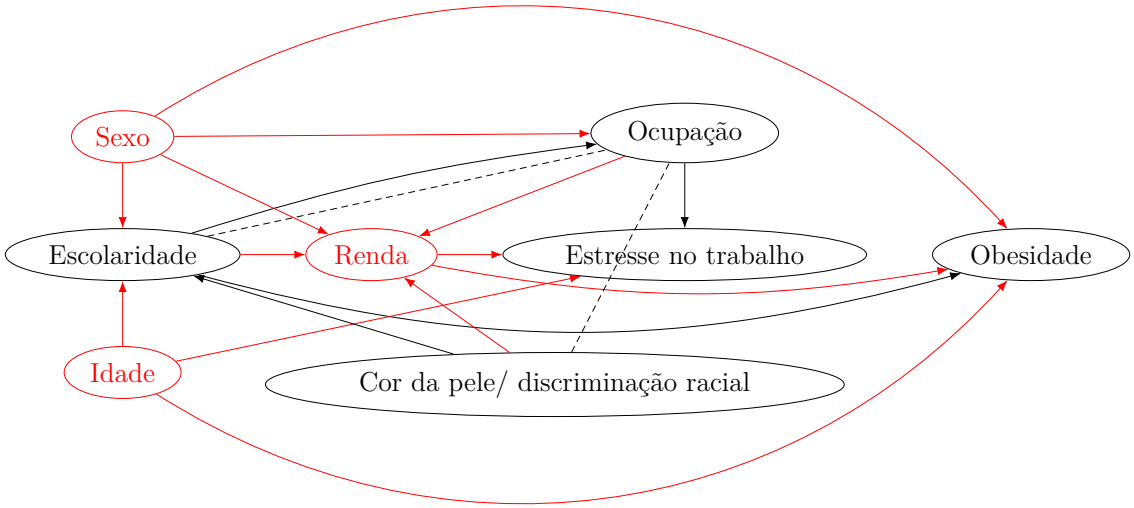

Figura 2.11: Em preto estão os caminhos abertos pelo critério *back-door*. Exemplo apresentado em Cortes et al. (2016).

Então, por (iii) verificamos pela Figura 2.11 que o conjunto $S = \{$`sexo, renda` e `idade`$\}$ não é suficiente para controlar o confundimento, já que caminhos que ligam a exposição ao desfecho permanecem abertos. Para que o conjunto de ajuste seja suficiente, pode-se incluir as variáveis `escolaridade` e `ocupação`, visto que todos os caminhos abertos incluem esses vértices.

2.4 Probabilidade e DAGs

Na análise de causa e efeito, fazemos o uso dos DAGs e o uso de probabilidades condicionais de eventos independentes para conectar a estrutura causal e os dados observados. A relação entre causalidade, que é indicada pelo DAG, e probabilidades é dada pela condição de Markov (Cortes et al., 2016).

Em Wasserman (2004) e Cortes et al. (2016) a condição de Markov é definida da seguinte forma: seja G um DAG com vertices $V = (X_1, \cdots, X_k)$, e $\mathbb{P}$ é uma distribuição de probabilidade para V com função de probabilidade f. G e $\mathbb{P}$ satisfazem

a condição de Markov se o conjunto de variáveis de V são independentes dos seus não descendentes (ou seja, não efeitos em G), dado o conjunto de seus pais (ou seja, causas diretas em G)

$$f(\upsilon) = \prod_{i=1}^{k} f(x_i \mid \pi_i),$$

onde π_i são os pais de X_i. O conjunto de distribuições representado por G pode ser denotado por $M(G)$.

Exemplo 2.4.1. *Para o DAG da Figura 2.10 a função de probabilidade f é fatorada da seguinte forma:*

$$\begin{aligned} f(&\text{estresse no trabalho, obesidade, renda, escolaridade, ocupação,} \\ &\text{cor da pele e ou discriminação racial, sexo, idade}) = \\ &f(\text{obesidade} \mid \text{estresse no trabalho, idade, sexo, escolaridade}) \\ &\times f(\text{estresse no trabalho} \mid \text{idade, renda, ocupação}) \\ &\times f(\text{renda} \mid \text{sexo, escolaridade, ocupação, cor da pele e ou discriminação racial}) \\ &\times f(\text{escolaridade} \mid \text{idade, sexo, cor da pele e ou discriminação racial}) \\ &\times f(\text{ocupação} \mid \text{sexo, escolaridade}) \\ &\times f(\text{cor da pele e ou discriminação racial}) \times f(\text{sexo}) \times f(\text{idade}) \end{aligned}$$

Logo podemos observar, por exemplo, que a função densidade de `estresse no trabalho` depende de `idade, renda` e `ocupação`, ou seja, seus pais. Mas não depende de `sexo, escolaridade` e `cor da pele e ou discriminação racial` que são seus avós. Isso se dá pela condição de Markov, que a variável é independente de seus ancestrais dado seus pais. Ou seja, uma vez que condicionada ao ancestral mais próximo pode-se descartar a informação dos ancestrais de seus pais.

O teorema apresentado por Wasserman (2004) mostra que $\mathbb{P} \in M(G)$ se, e somente se, a condição de Markov é mantida. Pode-se dizer que a condição é satisfeita quando cada variável W é independende de todas as causas ancestrais, dado seus pais.

Teorema 2.4.1. *Uma distribuição $\mathbb{P} \in M(G)$ se, e somente se, para cada variável W, a seguinte condição de Markov é válida:*

$$W \perp \widetilde{W} \mid \pi_W \tag{2.1}$$

onde $\widetilde{W}$ denota todas as outras variáveis exceto os pais e descendentes de W.

Exemplo 2.4.2. *Considere novamente o DAG apresentado em Cortes et al. (2016) de um estudo epidemiológico. Na Figura 2.12 podemos observar a relação apresentada pelo Teorema 2.4.1. A variável* `estresse no trabalho` *foi definida como* W*, logo as variáveis* `idade`*,* `renda` *e* `ocupação` *são definidas como* π_w*, os pais da variável* `estresse no trabalho`*. Por fim,* $\widetilde{W}$ *é definida pelas variáveis* `sexo`*,* `escolaridade` *e* `cor da pele/ discriminação racial`*, ou seja, todas as variáveis exceto os pais e descendentes de* W*.*

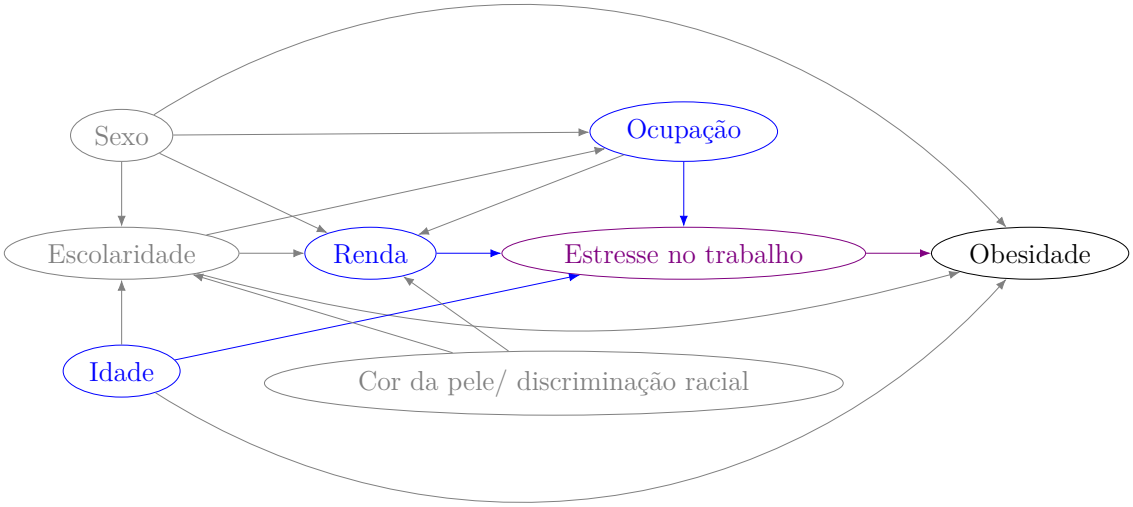

Figura 2.12: Exemplo 2.4.2. Em roxo está a variável que denota W; em azul as variáveis que representam π_W; e em cinza as variáveis que denotam $\widetilde{W}$.

2.5 Relações de Independência

A condição de Markov nos permite verificar as relações de independência condicional geradas por um DAG (Wasserman, 2004). Um modo mais fácil de verificar como a estrutura causal está relacionada com os dados através das relações de independência (a avaliação de consistência *DAG-dataset*), é feito por meio do critério de *d-separação* (Cortes et al., 2016).

Como descrito em Wasserman (2004), levando em consideração os DAGs das Figuras 2.13 e 2.14 as seguintes regras constituem o critério *d-separação*:

1. Quando Y não for um colisor, X e Z são d-conectadas, mas X e Z são d-separadas dado Y (primeiro e segundo DAG no topo).

2. Se X e Z colidem em Y, então elas são d-separadas, mas X e Z são d-conectadas dado Y (primeiro e segundo DAG na base).

3. O efeito do condicionamento no descendente de um colisor é o mesmo que do condicionamento no próprio colisor. Logo, na Figura 2.14 , X e Z são d-separadas, mas são d-conectadas dado W.

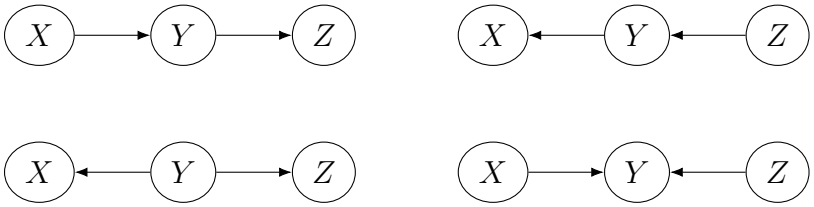

Figura 2.13: Apresentação de 4 DAGs; o quarto DAG possui um colisor, Y.

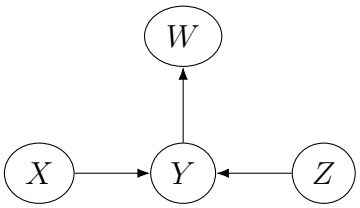

Figura 2.14: DAG com um descendente de um colisor.

Teorema 2.5.1. *Sejam X, Y e Z vértices disjuntos. Então $X \perp Y \mid Z$ se, e somente se, X e Y forem d-separados dado Z.*

Exemplo 2.5.1. *Considere o DAG da Figura 2.10. Pelo Teorema 2.5.1 podemos concluir que*

Estresse no trabalho ⊥ escolaridade | idade, renda, ocupação

Estresse no trabalho ⊥ sexo | idade, renda, ocupação

Estresse no trabalho ⊥ cor da pele e ou discriminação racial | idade, renda, ocupação

Obesidade ⊥ cor da pele e ou discriminação racial | estresse no trabalho, escolaridade, idade, renda, sexo

Obesidade ⊥ cor da pele e ou discriminação racial | ocupação, escolaridade, idade, renda, sexo

Obesidade ⊥ ocupação | estresse no trabalho, escolaridade, idade, renda, sexo

Ocupação ⊥ cor da pele e ou discriminação racial | escolaridade, sexo

Ocupação ⊥ idade | escolaridade, sexo

Renda ⊥ idade | escolaridade, sexo, cor da pele e ou discriminação racial

Ou seja, `estresse no trabalho` é d-separado de `escolaridade` dado `idade`, `renda` e `ocupação`. Isso segue para todas as implicações apresentadas no Exemplo 2.5.1.

3 Consistência DAG-dataset

Nem todo DAG será consistente com o dados gerados pelo processo causal. Por exemplo, se levarmos em consideração que cor e raça podem levar a um aumento do estresse no trabalho por meio de discriminação racial, incluiríamos uma aresta partindo de `Cor da pele e ou discriminação racial` para `estresse no trabalho` na Figura 2.10, porém, alguns autores indicam limitações em ralação a variável `cor da pele` como marcadora para discriminação racial (Cortes et al., 2016; Krieger et al., 1998). Assim, a escolha inadequada de uma variável ou do estabelecimento das relações causais entre variáveis, podem ter implicações para a validade do modelo de estudo (Cortes et al., 2016). Por isso a avaliação de consistência *DAG-dataset* torna-se imprescindível quando estamos fazendo uso de modelos causais. Muitos modelos causais não são simples de se analisar, a grande maioria deles terão mais de um único caminho entre as variáveis, então utilizamos o critério de *d-separação*, que pode ser aplicado em DAGs com alta complexibilidade, com o propósito de prever dependências que são compartilhadas por todos os conjuntos de dados gerados por aquele DAG (Pearl et al., 2016).

A aplicação do critério de *d-separação* pode acabar tornando-se uma tarefa difícil quando a estrutura do DAG contém um grande número de vértices e arestas (Cortes et al., 2016). Nessas situações, é recomendado o uso de softwares que permitam verificar as implicações estatísticas do modelo causal, como por exemplo o *DAGitty* (Textor et al., 2016).

As implicações estatísticas geradas pelo critério de *d-separação* podem ser avaliadas por alguma classe de testes de independência condicional para saber se são consistentes com os dados gerados pelo processo causal (Cortes et al., 2016). Se as implicações testáveis não são validadas pelos dados, talvez o DAG não seja uma representação precisa do processo causal que gera os dados (Cortes et al., 2016). Mas mesmo quando temos a suposição de consistência *DAG-dataset* atendida, não pode-se confirmar que a estrutura causal está correta, ainda deve-se considerar que um mesmo conjunto de relações de independência condicional é consistente com múltiplos grafos (Cortes et al., 2016), ou seja, todo o conjunto de DAGs possíveis

de se construir com um certo conjunto de variáveis é estatisticamente equivalente.

Neste capítulo será discutido mais a fundo sobre a avaliação de consistência *DAG-dataset*, quais as implicações testáveis e abordagens para testar essas implicações; e uma breve explicação sobre classes equivalentes.

3.1 Avaliação de consistência DAG-dataset

A avaliação de consistência *DAG-dataset* nos permite indentificar se os dados gerados pelo processo causal estão adequados ao modelo causal estabelecido pelo DAG. Essa avaliação é feita através das restrições estatisticamente testáveis, que são consequência do critério de *d-separação*, e podem ser encontradas em um DAG que contenha caminhos bloqueados entre duas variáveis (Textor et al., 2016). Então, por exemplo, o DAG da Figura 2.2 implica através do critério de *d-separação* que X e Z devem ser condicionalmente independentes dado Y ($X \perp Z \mid Y$), ou seja, X e Z são *d-separadas* dado Y. Ao testar essas implicações é possível verificar se o DAG é consistente com o conjunto de dados observados (Textor et al., 2016).

Dado o critério de *d-separação*, testamos se as variáveis são condicionalmente independentes para avaliar a consistência *DAG-dataset*, se as variáveis não são condicionalmente independentes, significa que o processo causal descrito pelo DAG não pode ter gerado os dados (Textor et al., 2016). Se as variáveis são condicionalmente independentes, se dará credibilidade as hipóteses do *DAG-dataset*, mas não prova que ele está correto (Textor et al., 2016).

Shipley (2002) apresenta algumas abordagens para se testar essas implicações estatísticas. Uma delas é por meio de uma regressão linear, que é equivalente a um teste de correlação parcial. Ou seja, regredimos X e Z em Y e depois testamos se a correlação entre os resíduos é igual à zero (ou seja, $H_0 = \rho_{X,Z|Y} = 0$). Se rejeitamos a hipótese nula, concluímos não haver consistência *DAG-dataset*, caso contrário, não há evidências de que o DAG e *dataset* não sejam consistentes. Se X e Z forem *d-separadas* por um conjunto de variáveis $\mathbf{Q}$={I,J,K,$\cdots$}, então deve-se regredir X e Z no conjunto de variáveis $\mathbf{B}$, possivelmente usando regressão não-linear, em que mesmo funções complicadas podem ser aproximadas por funções lineares, quadráticas ou cúbicas mais simples, pela vizinhança de um dado valor de X, e então realizar um teste para independência dos resíduos.

Para escolhermos o teste adequado, devemos olhar para distribuição dos resíduos, se estes forem normalmente distribuídos e linearmente relacionados, então pode-se usar o teste de correlação parcial de Pearson (Shipley, 2002). Se os resíduos parecem ter um comportamento monotônico, então pode-se usar o teste de correlação parcial de Spearman, que é muito parecido com o teste de correlação parcial de Pearson, só que neste as variáveis tem relação monotônica e a correlação parcial é aplicada às

classificações (*ranks*) dessas variáveis (Shipley, 2002). Já se os resíduos possuírem um comportamento mais complicado, então pode-se usar técnicas paramétricas de suavização, seguida de um teste de permutação (Shipley, 2002).

Os testes de independência condicional podem ser realizados com o pacote `dagitty`, porém, somente quando os resíduos forem normalmente distribuídos. O pacote comporta regressão linear e regressão polinomial local para o cálculo dos resíduos, suportando testes paramétricos e semi-paramétricos de independência condicional (Textor et al., 2016). No entanto, uma limitação do pacote `dagitty` é que ele ainda não possui em sua implementação testes não-paramétricos, não sendo possível assim, trabalhar com dados não-normais, ou seja, quando a correlação parcial é diferente de zero, mesmo quando as variáveis são condicionalmente independentes (Textor et al., 2016).

O coeficiente de correlação parcial de Pearson podem ser obtidos através da correlação parcial entre X e Y:

$$r_{X,Y|\boldsymbol{Q}} = \frac{-c_{XY}}{\sqrt{c_{XX} \times c_{YY}}},$$

onde, $-c_{XY}$ é a estimação da covariância parcial entre as variáveis X e Y. Shipley (2002) indica que a covariância parcial $-c_{XY}$ é obtida a partir de uma matriz de covariância de uma amostra M, em que o inverso desta matriz M é chamado de matriz de concentração C. O negativo dos elementos fora da diagonal c_{XY} fornece a covariância parcial entre as variáveis X e Y, condicionadas a um conjunto de variáveis $\mathbf{B}$ presente na matriz.

E através da correlação parcial de ordem n, que se dá entre duas variáveis condicionadas em outras n variáveis do conjunto $\mathbf{B}$. Por exemplo, a correlação parcial de ordem 1 entre X e Y condicionados em K:

$$r_{X,Y|K} = \frac{r_{XY} - r_{XK}r_{YK}}{\sqrt{(1 - r_{XK}^2)(1 - r_{YK}^2)}}.$$

O mesmo método pode ser usado para obter as correlações parciais de Spearman, a diferença é que primeiro as variáveis são ranqueadas, então em seguida pode-se aplicar o mesmo procedimento que para as correlações parciais de Pearson (Shipley, 2002).

3.2 Implicações testáveis

Na seção anterior foi citado que os modelos causais tem implicações testáveis no conjunto de dados gerado pelo processo causal. Ou seja, as implicações testáveis do modelo dizem respeito as independências condicionais estabelecidas entre as

variáveis, obtidas através do critério *d-separação*.

Então por exemplo, se tivermos um grafo acíclico dirigido G (Figura 2.2) que possa ter gerado o conjunto de dados D, o critério de *d-separação* irá nos dizer quais variáveis em G são condicionalmente independentes. Se listarmos as implicações testáveis do critério de *d-separação* em G, iremos observar que X e Z devem ser condicionalmente independentes em Y. Então, vamos supor que quando estimamos as probabilidades com base em D, os dados sugerem que X e Z não são condicionalmente independêntes em Y. Logo, rejeitamos G como possível modelo causal para D (Pearl et al., 2016).

Utilizando o exemplo de (Cortes et al., 2016) (Figura 2.10). Considerando a relação `estresse no trabalho` $\perp$ `escolaridade` | `renda`, sabemos que `estresse no trabalho` e `escolaridade` são *d-separados* dado `renda`. Então se regredimos `estresse no trabalho` em `renda` e `escolaridade` iremos obter a reta que melhor se ajusta aos dados (Pearl et al., 2016):

$$estresse\ no\ trabalho = r_{renda}\ renda + r_{escolaridade}\ escolaridade$$

Assim, se $r_{escolaridade}$ (correlação entre os resíduos) não for igual à zero, saberemos que `estresse no trabalho` e `escolaridade` não são condicionalmente independentes dado `renda`, logo, o modelo não está correto (Pearl et al., 2016).

Textor et al. (2016) aponta algumas causas que devem ser consideradas quando implicações do tipo $X \perp Z \mid Y$ são inconsistentes com o conjunto de dados gerado pelo processo causal, como:

(i) especificação incorreta das relações entre as variáveis observadas incluídas no DAG;

(ii) omissão de uma variável latente no DAG que é causa comum de duas ou mais variáveis;

(iii) erro de medição em uma ou mais variáveis incluídas no DAG, sendo assim, é possível que exista multicolinearidade entre as variáveis preditoras.

Essas possíveis causas requerem uma maior consideração pelos conhecimentos prévios acerca das relações entre as variáveis contidas no DAG (Textor et al., 2016). Uma forma de verificar essas causas é olhando para as implicações testáveis que foram inconsistentes com o conjunto de dados, tentando estabeler relações para essas inconsistências com base no conhecimento prévio sobre o assunto, assim verificando quais vértices ou arestas devam ser adicionados ou retirados do modelo causal.

3.3 Classes equivalentes DAGs

As classes equivalentes dos DAGs são baseadas nas implicações testáveis, logo, não dependem das exposições nem desfechos aos quais o DAG foi originalmente pensado (Textor et al., 2016). Então, mesmo quando temos a suposição de consitência *DAG-dataset* atendida, não prova que o DAG está correto, e uma das limitações é que diferentes DAGs podem apresentar as mesmas implicações testáveis (Textor et al., 2016).

Podemos observar isso com os três primeiros DAGs apresentados da Figura 2.13. Todos eles tem a mesma implicação testável $X \perp Z \mid Y$, mesmo com diagramas causais diferentes. O segundo DAG da Figura 2.13 (da esquerda para a direita no topo) é simetricamente o oposto do primeiro DAG da Figura 2.13 (da esquerda para direita no topo), e o primeiro DAG da Figura 2.13 (da esquerda para a direita na base) não possuí caminho causal entre X e Z, mas ambas as variáveis causam Y. Esses três DAGs implicam que X e Z são condicionalmente independentes dado Y, mesmo possuindo interpretações causais diferentes.

Apesar das classes de equivalência não precisarem levar em consideração a exposição e o desfecho, elas possuem uma grande relação com os conjuntos de ajustes suficientes (que dependem da exposição e desfecho), pois se o mesmo conjunto de ajustes suficiente se aplica a relação entre exposição e desfecho em um conjunto de DAGs presente em uma classe equivalente, reforça que o conjunto de dados gerado pelo processo causal é válido (Textor et al., 2016).

4 O DAGitty

Quando um DAG possui muitas variáveis, é complicado verificar as relações de independência condicional sem o auxílio de alguma ferramenta computacional. Uma maneira de facilitar esse processo é utilizar o aplicativo da Web o *'DAGitty'* ou o pacote de mesmo nome disponível no software R, desenvolvidos por Textor et al. (2016), em que é possível desenhar, editar, testar e analisar os DAGs.

Neste capítulo, será apresentado uma breve introdução ao uso dessas ferramentas. Para ilustrar o uso das funções disponíveis no *'DAGitty'* será utilizado um exemplo de uma investigação epidemiológica apresentado em Kleinbaum et al. (1982), onde o objetivo era avaliar a suposta associação do nível de catecolaminas endógenas (`CAT`) no sangue com a incidência decorrente de doença cardíaca coronária (`CHD`). Há muitas evidências com base em estudos em animais e humanos de que os fatores psicossociais estimulam a medula adrenal para liberar das duas catecolaminas na corrente sanguínea, a norepinefrina (noradrenalina) e epinefrina (adrenalina). Porém, não há muitas informações de como os mecanismos biológicos vinculam o ambiente social de uma pessoa à ocorrência de doenças, ou sobre o papel que outros fatores de risco biológico desempenham neste processo, como por exemplo, o colesterol (Kleinbaum et al., 1982).

Os dados são refetentes ao estudo de coorte derivado do Estudo sobre Doenças Cardíacas do Condado de Evans (*Evans County Heart Disease Study*), realizado entre 1960 e 1969. Os dados dizem respeito a uma coorte de 609 indivíduos do sexo masculino, brancos, com idades entre 40 e 76 anos, livres de doença coronariana e residentes no condado de Evans, Geórgia, em 1960. Após sete anos, toda a coorte foi reexaminada, e 71 novos casos de doença coronariana foram identificados. Kleinbaum et al. (1982), para fins didáticos, geraram artificialmente os níveis de catecolaminas; todas as outras variáveis foram retiradas do estudo original do Condado de Evans, assim como, com a exceção de `CHD`, todas as variáveis foram medidas na linha de base, em 1960. Na Tabela 4.1 são apresentadas as definições de cada variável presente no conjunto de dados.

Na Tabela 4.2 é apresentado o resultado da análise descritiva da comparação de

Tabela 4.1: Descrição das variáveis do banco de dados do Condado de Evans

Nome	Descrição	Código
ID	Identificador de cada observação	Contagem
CHD	Incidência de doença cardíaca coronária	0 = ausência da doença 1 = presença da doença
CAT	Nível de catecolaminas séricas	0 = baixo 1 = alto
AGE	Idade	Anos
CHL	Colesterol sérico	mg/100 mL
SMK	Tabagismo	0 = nunca fumou 1 = já fumou
ECG	Anormalidade no eletrocardiograma	0 = ECG normal 1 = ECG com anormalidade
DBP	Pressão arterial diastólica	mmHg
SBP	Pressão arterial sistólica	mmHg
HPT	Pressão alta	0 = não tem pressão alta 1 = tem pressão alta

`CAT` com as outras variáveis presentes no conjunto de dados. A única variével que não apresenta nenhuma diferença entre os níveis de catecolaminas, foi o tabagismo (p-valor = 0.912). As variáveis que apresentam uma maior diferença entre os níveis de catecolaminas foram `CHD, ECG` e `HPT`.

Tabela 4.2: Análise descritiva da comparação de CAT com demais variáveis do conjunto de dados do Condado de Evans

		Nível baixo de CAT N = 487	**Nível alto de CAT** N = 122
CHD:			
	Não	443 (91.0%)	95 (77.9%)
	Sim	44 (9.03%)	27 (22.1%)
AGE		51.9 (8.60)	61.1 (8.12)
CHL		215 (40.1)	199 (36.4)
SMK:			
	Nunca fumou	177 (36.3%)	45 (36.9%)
	Já fumou	310 (63.7%)	77 (63.1%)
ECG:			
	ECG normal	396 (81.3%)	47 (38.5%)
	ECG alterado	91 (18.7%)	75 (61.5%)
DBP		88.7 (12.7)	101 (16.9)
SBP		138 (20.1)	177 (30.8)
HPT:			
	Normal	330 (67.8%)	24 (19.7%)
	Alta	157 (32.2%)	98 (80.3%)

Já na Figura 4.1, temos as comparações entre todas as variáveis presentes no conjunto de dados, então por exemplo, podemos observar que as variáveis que estão mais correlacionadas são `DBP` e `SBP`, com uma correlação positiva de 0,753. As variáveis menos correlacionadas são `CHL` e `AGE`, com uma correlação negativa de -0,004, indicando que o nível de colesterol está pouco relacionado com a idade do indivíduo, ainda, que indivíduos com nível baixo de catecolamina tem um nível de colesterol mais alto em pessoas mais jovens, já indivíduos com nível alto de catecolaminas tem um nível de colesterol mais alto em pessoas mais velhas.

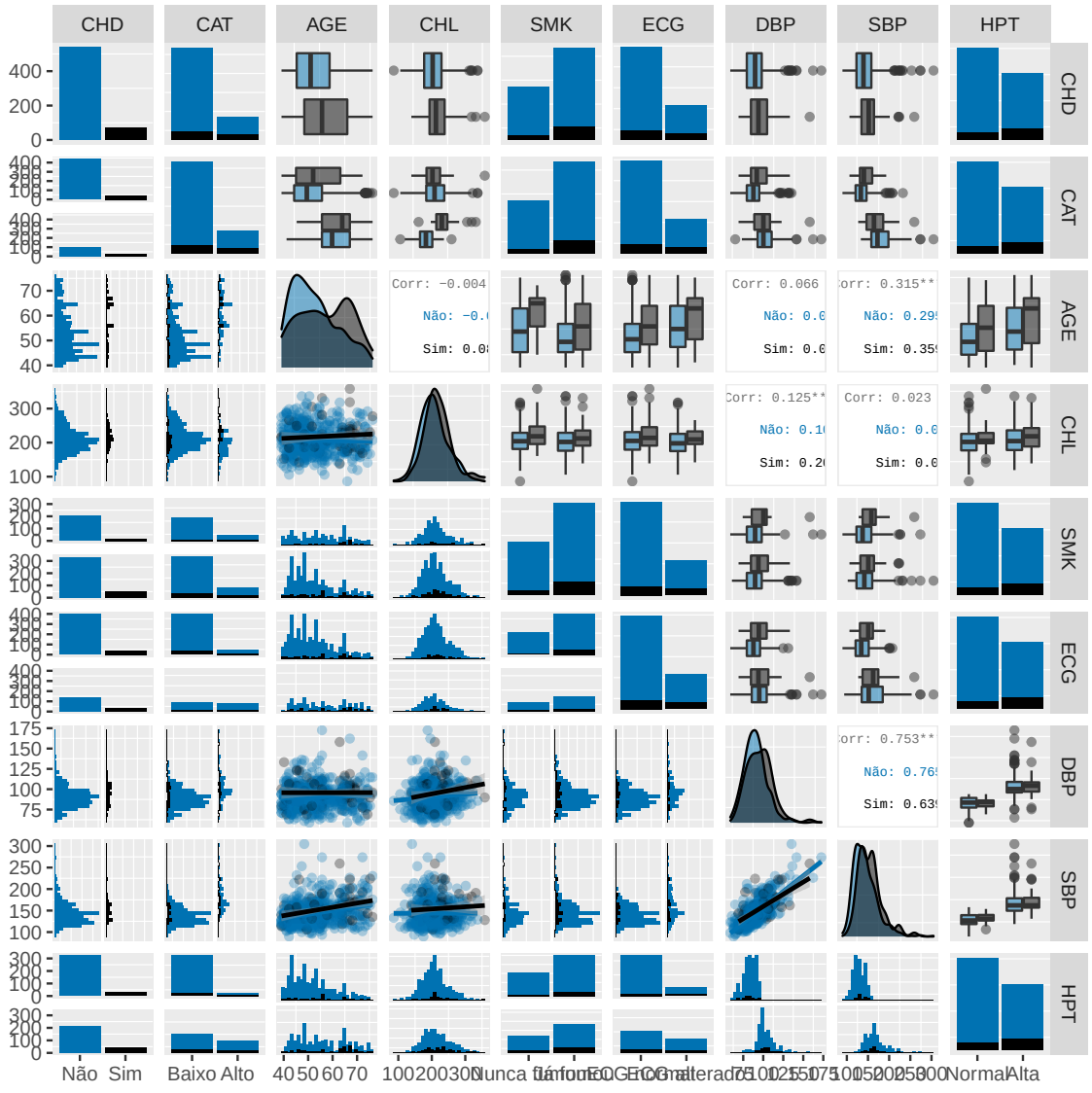

Figura 4.1: Análise descritiva da comparação entre as variáveis do conjunto de dados do Condado de Evans.

4.1 DAGitty Web

O *'DAGitty'* Web, disponível para uso online em dagitty.net, é uma ferramenta muito intuitiva, sendo assim, de uso mais simples e fácil que o pacote `dagitty`.

Após acessar o dagitty.net e em sua página inicial clicar em *'Launch DAGitty online in your browser'* (Inicie o DAGitty online em seu navegador), o usuário é direcionado para a página (Figura 4.2) onde é possível desenhar e analisar as relações de independência condicional estabelecidas pelas variáveis. A página é dividida em três áreas, a central é onde o DAG será desenhado ou editado, também são apresentadas opções de modelo (referente ao desenho do DAG), exemplos, como construir um DAG no *DAGitty*, layout e ajuda.

Na área à esquerda temos as opções de estilo do DAG, como por exemplo em *'View mode'* (Modo de visualização), é permitido escolher o tipo de DAG derivado com que deseja trabalhar, ou seja, o DAG que foi transformado para se adequar ao propósito do estudo; ou *'Effect analysis'* (Análises de efeito), com a função *'Atomic direct effects'* (Efeitos diretos atômicos) que identifica os caminhos causais diretos, uma opção interessante quando se está trabalhando com DAGs muito grandes, consequentemente produzindo muitas setas que podem gerar caminhos indiretos. Também na área a esquerda é apresentada a função *'Diagram style'* (Estilo do diagrama) onde é possível escolher a opção *'Classic'* (Clássico) em que os vértices e seus rótulos são separados, ou a opção *'SEM-like'* (Modelagem da Equação Estrutural) em que os rótulos estão dentro dos vértices; em *'Coloring'* (Coloração) é possivel escolher destacar com diferentes cores os caminhos abertos e fechados e em *´Legend'* (Legenda) o que cada cor de vértice e aresta significam.

Por fim, na área à direita temos a identificação do efeito causal (*'Causal effect identification'*), com três opções possíveis: conjunto de ajustes (de efeito total ou efeito direto) ou variáveis intrumentais, esta última é muito utilizada quando os fatores de confusão não são observados (não mensurados), não sendo possível estimar o efeito causal por um simples ajuste de covariáveis. Nesta área são apresentadas também as implicações testáveis, o código do modelo e o resumo das variáveis presentes no DAG.

Na Figura 4.3 são apresentadas as opções disponíveis no menu da área central do *'DAGitty'*. Em *'Model'* são apresentadas as opções de criar um novo modelo (ou seja, desenhar um novo DAG), publicar o DAG no dagitty.net, e a partir disso, é possível carregar, editar e excluir no *'DAGitty'* Web os DAGs uma vez já publicados. Também é possível exportar em diferentes formatos os DAGs desenhados, como PDF, PNG, JPEG, SVG e em código LaTeX. Em *'Exemples'* são apresentados 11 exemplos didáticos, alguns deles retirados de artigos científicos. Já em *'How to...'* estão disponíveis breves explicações de como construir um DAG no *'DAGitty'* Web,

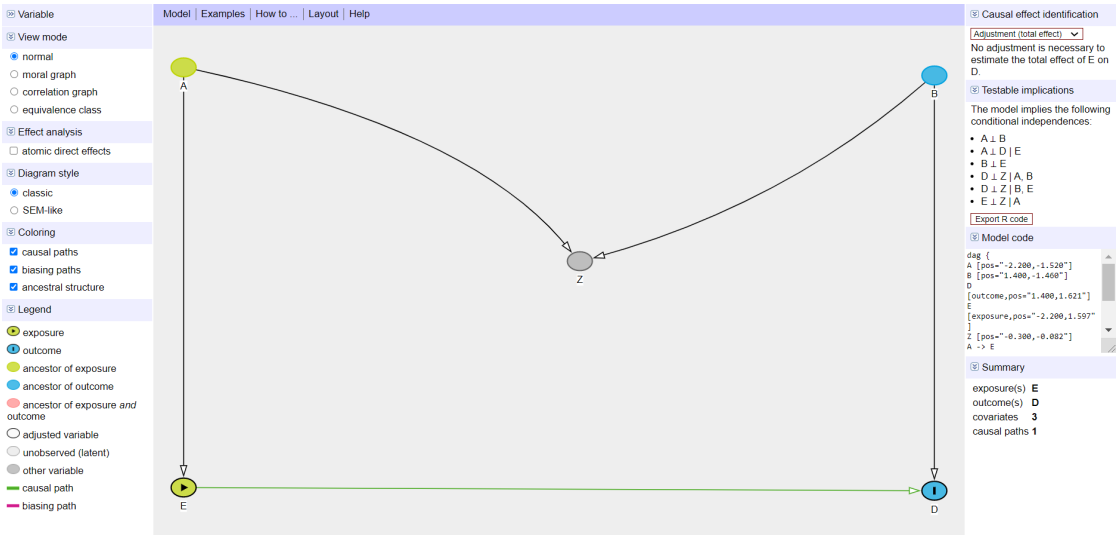

Figura 4.2: Página inicial do DAGitty.

como por exemplo como adicionar uma nova variável ao DAG (que é apenas dar um duplo-clique na tela onde o DAG está sendo desenhado), ou como adicionar uma aresta (clicar primeiro no vértice onde começa a aresta e depois no vértice onde termina). Em *'Layout'* está disponível a função em que o aplicativo gera automaticamente o layout do DAG desenhado. Por último temos o *'Help'* com o manual de uso do *'DAGitty'* e alguns outros materiais complementares sobre Grafos Acíclicos Dirigidos e suas implicações.

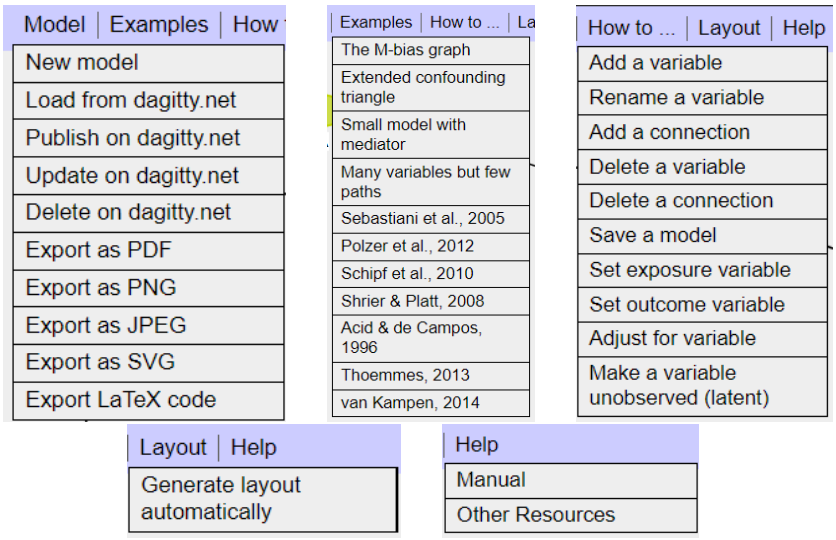

Figura 4.3: Opções disponíveis na área central do DAGitty.

Podemos desenhar o DAG do exemplo apresentado em Kleinbaum et al. (1982) através do *'DAGitty'* Web. Então na Figura 4.4 podemos observar em verde os caminhos causais e em rosa os caminhos com viés de confundimento. Na área à direita é apresentado o conjunto de ajustes suficientes S para estimar o efeito causal

de `CAT` (exposição) em `CHD` (desfecho): S= {`AGE, SMK`}. O conjunto de ajustes suficiente S é obtido pela aplicação interna do critério *back-door*, apresentado na Seção 2.3. As relações de independência condicional entre as variáveis também podem ser vistas nessa área.

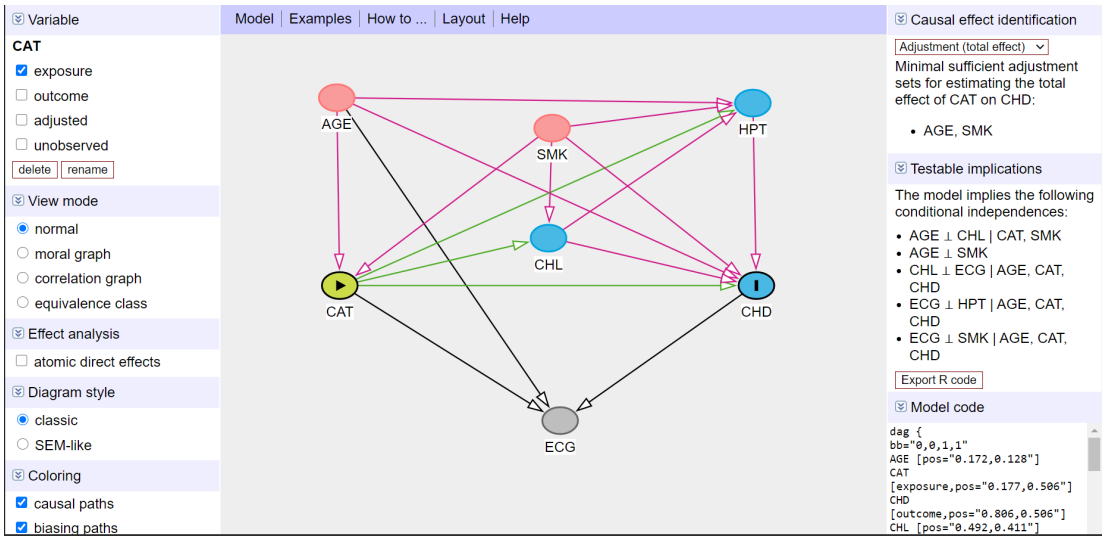

Figura 4.4: DAG do exemplo apresentado em Kleinbaum et al. (1982) feito no dagitty.net.

Este DAG não é apresentado em Kleinbaum et al. (1982), mas foi construído com base em sugestões obtidas na literatura, porém não chega a configurar um modelo causal teórico. Nosso objetivo com a apresentação do DAG da Figura 4.4 era prioritariamente ilustrar os métodos discutidos neste trabalho.

4.2 O pacote dagitty

Agora que o *'DAGitty'* Web para o uso online já foi brevemente apresentado, vamos verificar como podemos utilizar o pacote `dagitty` disponível para uso no software R.

Com o pacote `dagitty` podemos trabalhar com o conjunto de dados gerados pelo processo causal, sendo assim, possível fazer a avaliação de consistência *DAG-dataset* (dada normalidade dos dados), verificar as classes equivalentes, além das funções já apresentadas no *'DAGitty'* Web como o conjunto de ajustes suficientes e as implicações testáveis.

O primeiro passo a ser feito, é a instalação do pacote `dagitty` no software R, em seguida pode ser feita a leitura dos dados do exemplo e o *download* do DAG desenhado anteriormente no *DAGitty* Web, assim, não sendo preciso desenhar o DAG diretamente no software R.

```
#Instalar o pacote dagitty
install.packages("dagitty")

#Carregar o pacote dagitty
library(dagitty)

#Leitura dos dados
evans <- read.table("evans.txt", head =T)

#Carregar o DAG a partir do DAGitty web
dag <- downloadGraph("dagitty.net/m--wcZT")

#Desenhar o DAG
plot(dag)
```

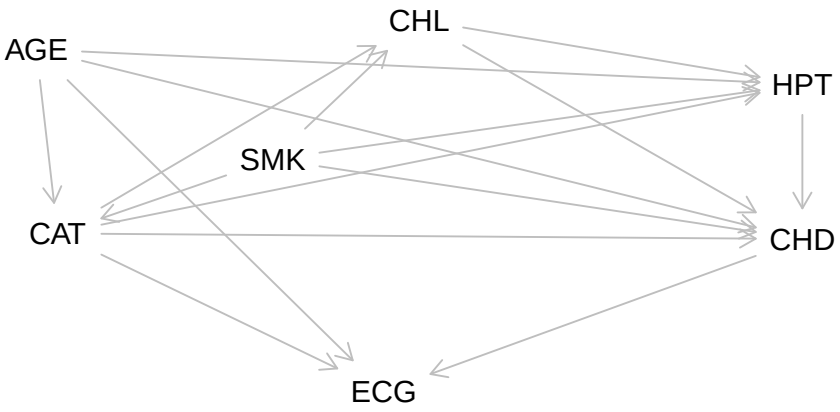

Figura 4.5: Plot do DAG do exemplo apresentado em Kleinbaum et al. (1982) (Figura 4.4) carregado a partir do *'DAGitty'* web.

A avaliação de consistência *DAG-dataset* pode ser feita através da função `localTests` que deriva as implicações testáveis do DAG e as testa em relação ao conjunto de dados, ou seja, a função `localTests` aplica o critério de *d-separação* no DAG enumerando as relações de independência condicional implícitas no mesmo, e em seguida, é realizado um teste de correlação parcial (apresentado na Seção 3.1) para cada uma das independências condicionais observadas. Os resultados obtidos através da função `localTests` podem ser vistos na Tabela 4.3.

```
#Criterio d-separacao

d.sep <- localTests(dag, evans)
```

Tabela 4.3: Estimativas pontuais intervalares e p-valor das correlações parciais entre as variáveis do DAG "CAT-CHD"presentes nas implicações testáveis.

	Estimativa	**p-valor**	**2.5%**	**97.5%**
AGE ⊥ CHL | CAT, SMK	0.062	0.1253	-0.017	0.141
AGE⊥ SMK	-0.139	0.0006	-0.216	-0.060
CHL ⊥ ECG | AGE, CAT, CHD	0.000	1.0000	-0.080	0.080
ECG ⊥ HPT | AGE, CAT, CHD	0.104	0.0104	0.025	0.182
ECG ⊥ SMK | AGE, CAT, CHD	-0.027	0.5057	-0.107	0.053

Com base no resultado obtido através da função `localTests` do pacote `dagitty`, podemos concluir que o DAG "CAT-CHD"não é consistente com o *dataset*, visto que as implicações testáveis para mais de um par de variáveis foram rejeitadas pela hipótese nula, de que a correlação parcial entre os resíduos é igual à zero. Isso pode ter ocorrido pela especificação incorreta das relações dessas variáveis, assim como a omissão de alguma variável latente que é causa comum desses pares de variáveis, possivelmente existindo multicolinearidade entre as variáveis preditoras. Podemos observar que as implicações testáveis que foram inconsistentes com o conjunto de dados $AGE \perp SMK$ e $ECG \perp HPT \mid AGE, CAT, CHD$, parecem ter uma relação causal que não foi especificada no modelo, como a idade (`AGE`) está relacionada com ser ou não fumante (`SMK`), assim como possuir ou não uma anormalidade no eletrocardiograma (`ECG` pode estar relacionado com ter ou não pressão alta (`HPT`). Então se essas relações entre as variáveis existirem e foram omitidas no modelo, podemos adicinar a aresta entre esses vértices, assim estabelecendo a relação causal entre essas variáveis, e teremos que o *DAG-dataset* é consistente, já que não teremos nenhuma implicação rejeitada pela hipótese nula.

Quando trabalhamos com DAGs muito grandes, é possível que esses possuam muitas implicações testáveis. Uma maneira de contornar isso é corrigir os p-valores obtidos em testes múltiplos. Há vários métodos para correção do p-valor disponíveis no software R, como por exemplo o método de Holm-Bonferroni. O exemplo que está sendo trabalhado originalmente possuí cinco implicações testáveis, logo, não seria necessário fazer a correção dos p-valores. Porém, para fins didáticos será apresentado como essa correção é feita.

```
#Correcao de Holm-Bonferroni
d.sep$p.value <- p.adjust(d.sep$p.value)
```

Após realizada a correção de Holm-Bonferroni, filtramos os pares de variáveis que rejeitaram a hipótese nula ($H_0 : \rho = 0$), com um valor de corte arbitrário. Os pares de variáveis que o teste rejeitou a hipótese nula nos apontam para uma inconsistência entre o DAG e o *dataset*.

Tabela 4.4: Estimativas pontuais intervalares e p-valor das correlações parciais entre as variáveis do DAG "CAT-CHD"presentes nas implicações testáveis corrigido pelo método de Holm-Bonferroni.

	Estimativa	**p-valor**	**2.5%**	**97.5%**
AGE ⊥ CHL | CAT, SMK	0.062	0.3758	-0.017	0.141
AGE⊥ SMK	-0.139	0.0028	-0.216	-0.060
CHL ⊥ ECG | AGE, CAT, CHD	0.000	1.0000	-0.080	0.080
ECG ⊥ HPT | AGE, CAT, CHD	0.104	0.0415	0.025	0.182
ECG ⊥ SMK | AGE, CAT, CHD	-0.027	1.0000	-0.107	0.053

```
#Teste com p-valor inferior ao corte arbitrario

d.sep <- d.sep[d.sep$p.value<0.05,]
```

Tabela 4.5: Estimativas pontuais intervalares e p-valor das correlações parciais entre as variáveis do DAG "CAT-CHD"presentes nas implicações testáveis menores que o valor de corte arbitrário de 0.05.

	Estimativa	**p-valor**	**2.5%**	**97.5%**
AGE⊥ SMK	-0.139	0.0028	-0.216	-0.060
ECG ⊥ HPT | AGE, CAT, CHD	0.104	0.0415	0.025	0.182

A função `plotLocalTestResults` gera um gráfico (Figura 4.6) com os resultados obtidos a partir da função `localTests`, então para cada teste, os coeficientes de correlações parciais e seus intervalos de confiança são apresentados. Nesse caso, o gráfico foi gerado a partir dos resultados obtidos do teste com um corte arbitrário de $p - valor < 0,05$, após realizada a correção de Holm-Bonferroni.

```
#Grafico resumo

plotLocalTestResults(d.sep, xlab = "Correlacao Parcial",
    xlim = c(-0.4,0.4), ylim = c(0.5,7),frame = FALSE)
```

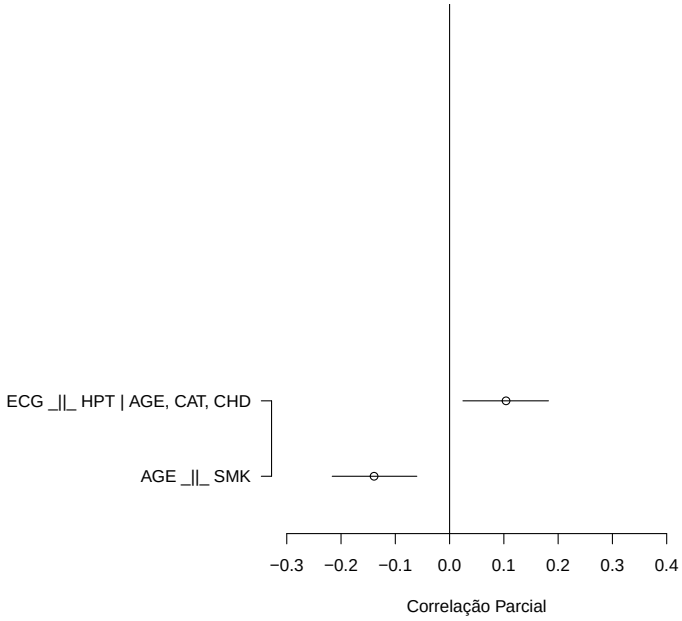

Figura 4.6: Correlações empíricas, cada uma relacionada as implicações testáveis separadas, para as quais o p-valor corrigido é menor que um valor de corte arbitrário de 0,05.

As estimativas das correlações empíricas podem ser tanto positivas quanto negativas. Porém, é tido como regra que quanto mais afastadas de zero forem as estimativas, mais inconsistentes são as implicações testáveis com o conjunto de dados (Textor et al., 2016).

É possível identificar conjuntos de ajustes mínimos suficientes e robustos combinando a avaliação da consistência *DAG-dataset* com a identificação de conjuntos de ajustes válidos para DAGs estatisticamente equivalentes.

O conjunto de ajustes mínimos suficientes pode ser obtido através da função `adjustmentSets`.

```
#Conjunto de ajustes suficientes

adjustmentSets(dag, "CAT", "CHD")
```

```
{AGE, SMK}
```

Podemos observar que a função `adjustmentSets` estimou o mesmo conjunto de ajuste mínimo suficiente (`{AGE, SMK}`) que a função Identificação do efeito causal (*'Causal effect identification'*) disponível no *'DAGitty'* Web. Ou seja, com ferramentas diferentes podemos obter os mesmos resultados.

A função `equivalenceClass` apresenta as classes de equivalência do DAG (Figura 4.7).

```
#Determinar as classes equivalentes

eq <- equivalenceClass(dag)
```

```
AGE -> CAT
AGE -> CHD
AGE -> ECG
AGE -> HPT
CAT -> CHD
CAT -> CHL
CAT -> ECG
CAT -> HPT
CHD -- HPT
CHD -> ECG
CHL -> CHD
CHL -> HPT
SMK -> CAT
SMK -> CHD
SMK -> CHL
SMK -> HPT
```

Figura 4.7: Classes de equivalência. Arestas que têm a mesma direção em todos os diferentes DAGs são exibidas normalmente, enquanto as arestas com direções diferentes nos diferentes DAGs são exibidos sem pontas de seta.

Existem 15 setas no DAG de `CAT` cuja direção é a mesma em toda a classe de equivalência. Portanto, se ocorresse um erro na especificação da direção de qualquer uma dessas setas, levaria a uma alteração nas implicações testáveis e, isso é potencialmente detectável usando a avaliação de consistência *DAG-dataset*.

```
#Conjunto de ajuste suficiente para toda classe equivalente

adjustmentSets(eq, "CAT", "CHD")
```

```
{AGE, SMK}
```

O conjunto de ajuste sufiente identificado para toda classe equivalente revela que o conjunto de ajuste suficiente original também é válido para todos os DAGs equivalentes.

O pacote `dagitty` possui ainda diversas funções que não foram utilizadas no exemplo apresentado, como por exemplo:

- `backDoorGraph`: remove as arestas originadas da exposição;
- `impliedConditionalIndependencies`: lista todas as relações de independência condicional presentes no modelo causal;
- `ancestorGraph`: gera um gráfico com apenas os vértices de interesse em **Q**, seus ancestrais e as arestas entre eles. Todos os outros vértices são descartados;

- `as.dagitty`: converte, se possível, um objeto a um formato DAGitty;
- `completeDAG`: gera um DAG completo a partir das variáveis fornecidas;
- `randomDAG`: gera um DAG aleatório com N variáveis;
- `getExemple`: oferece acesso aos exemplos disponíveis no *'DAGitty'* web.

4.3 O pacote ggdag

Há disponível o pacote `ggdag`, desenvolvido por Barrett (2021), que é uma extensão do pacote `dagitty` e pode ser usado em conjunto com o pacote `ggplot2`, de Wickham (2016). Este pacote possui uma gama maior de opções para *layout* que o pacote `dagitty`. Vamos reproduzir o exemplo 4.5, agora utilizando o pacote `ggdag`.

O primeiro passo é instalar o pacote `ggdag`, em seguida baixamos novamente o DAG a partir do *'DAGitty'* Web. A função ggdag, que utiliza como objeto de entrada o DAG construído no *'DAGitty'*, constrói o gráfico do DAG em layout similar aos gráficos gerados pelo pacote ggplot2. A função `ggdag_ancestors` indica no gráfico do DAG todos os ancestrais da variável de interesse. Já a função `ggdag_adjustument_set` desenha no gráfico do DAG os conjuntos de ajustes suficientes. As demais funções do pacote ggdag apresentam saídas no mesmo layout com as mesmas opções de ajustes gráficos, tais como vértices, textos e rótulos, como apresentado a seguir.

```
#Instale o pacote ggdag

install.packages("ggdag")

#Carregar o pacote

library(ggdag)

#Carregar o DAG a partir do DAGitty Web

dag2 <- downloadGraph("dagitty.net/m--wcZT")

#Desenhar o DAG

ggdag(dag2, text_col = "grey80", text_size = 2.7, node_size = 12)+
  theme_dag()+
  ggtitle("DAG - Catecolaminas")
```

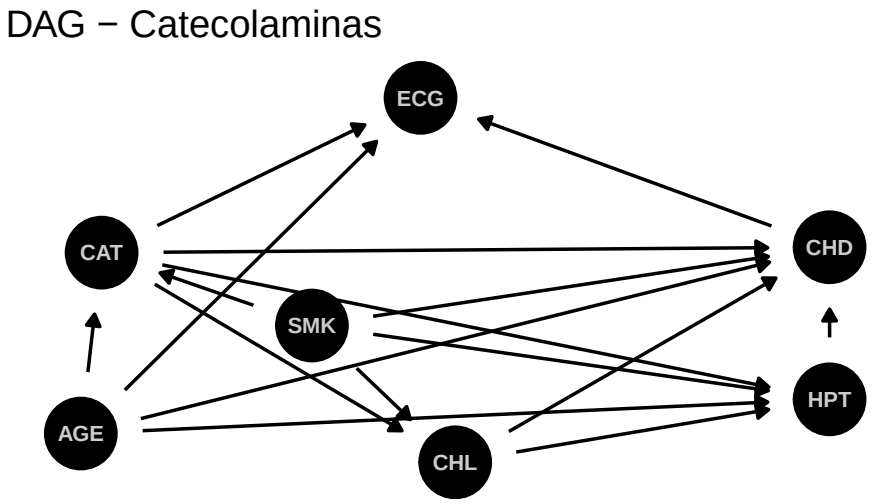

Figura 4.8: DAG do exemplo , reproduzido com o pacote `ggdag`.

Na função de plotagem do DAG , foram especificadas a cor do texto (`text_col`), o tamanho do texto (`text_size`) e tamanho do vértice (`node_size`).

Podemos verificar a relação familiar entre as variáveis graficamente; diversas funções no pacote `ggdag` auxiliam nesse objetivo, porém com diferentes propostas, como por exemplo, a função `ggdag_children` apresenta todos os filhos da variável de interesse, já a função `ggdag_parents` apresenta todos os pais da variável indicada.

```
#Relacao de ancestralidade

ggdag_children(dag2, "CAT", node_size = 12)+
  geom_dag_text(colour = "white", size = 4)+
  theme_dag()+
  scale_color_manual(values = c("#0072B2", "black"),
    na.value = "grey80")+
  ggtitle("Relacao de ancestralidade")
```

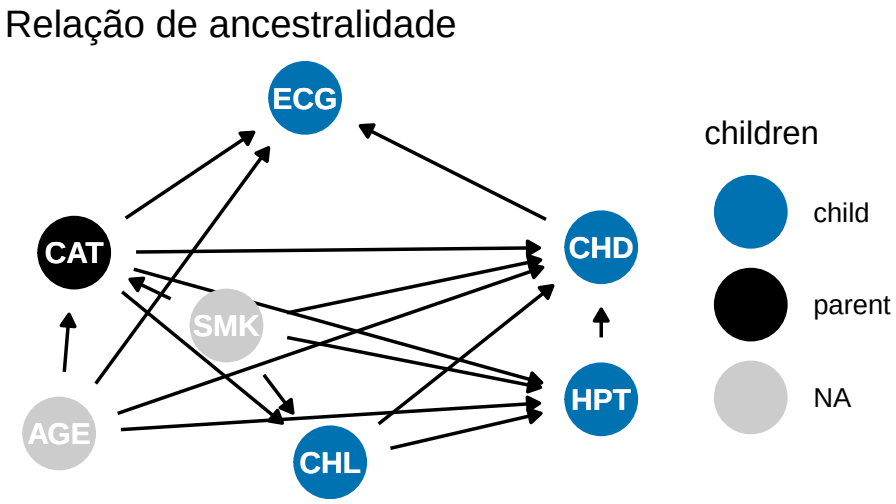

Figura 4.9: Relação de ancestralidade do DAG 4.8 apresentada pelo pacote `ggdag`.

O conjunto de ajustes mínimo suficiente (Figura 4.10), obtido anteriormente no *'DAGitty'* Web e pelo pacote `dagitty` com a função `adjustmentSets`, pode ser visualizado graficamente através da função `ggdag_adjustment_sets` do pacote `ggdag`.

```
#Conjunto de ajustes suficientes

ggdag_adjustment_set(dag2, node_size = 12)+
  geom_dag_text(colour = "white", size = 4)+
  theme_dag()+
  scale_color_manual(values = c("#0072B2", "black"),
    na.value = "grey80")+
  ggtitle("Conjunto de ajuste minimo suficiente")
```

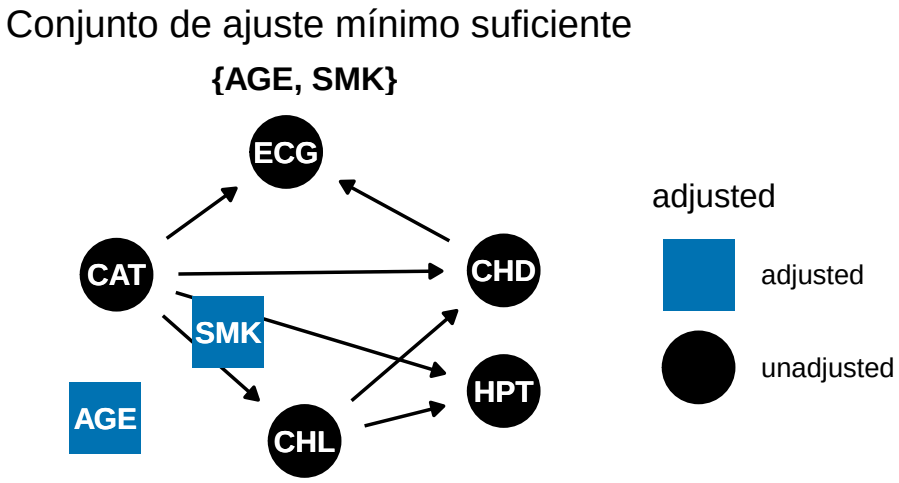

Figura 4.10: Conjunto de ajustes mínimo suficiente do DAG 4.8 apresentado pelo pacote `ggdag`.

Podemos visualizar todos os caminhos abertos (Figura 4.11) presentes no DAG através da função `ggdag_paths`, tornando assim, mais fácil a visualização de todos os caminhos que possuam alguma associação linear entre variáveis.

```
#Caminhos abertos

ggdag_paths(dag2, node_size = 10)+
  geom_dag_text(colour = "white", size = 4)+
  theme_dag()+
  scale_color_manual(values = c("#0072B2", "black"),
    na.value = "grey80")+
  ggtitle("Caminhos abertos de CAT para CHD")
```

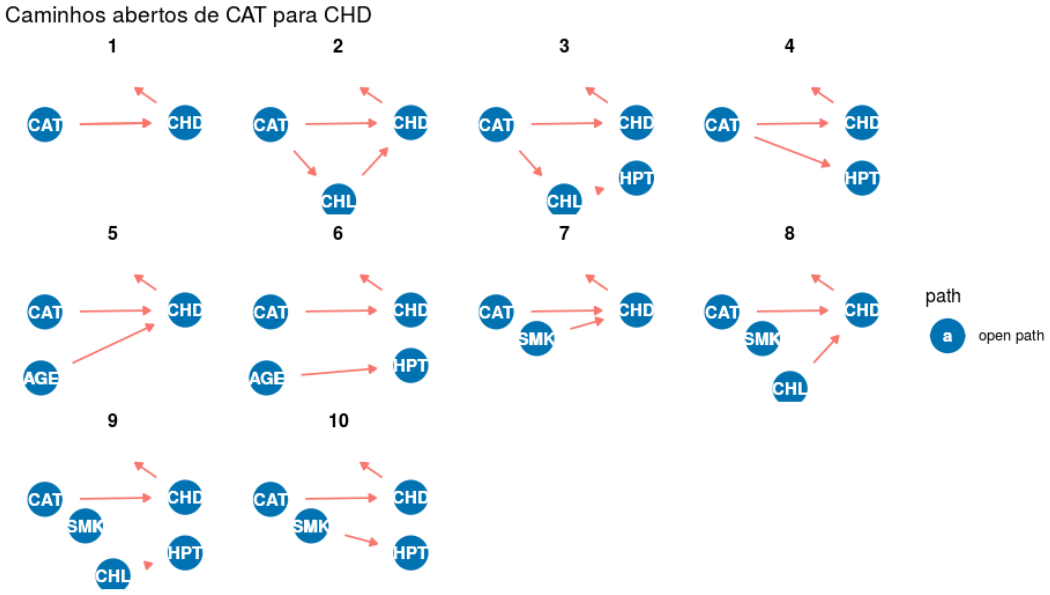

Figura 4.11: Caminhos abertos do DAG 4.8 apresentados pelo pacote `ggdag`.

O pacote `ggdag` nos fornece basicamente as mesmas funções que o pacote `dagitty`, por exemplo, o `ggdag` não possui uma função para avaliar a consistência *DAG-dataset*, porém possui mais opções de possíveis adaptações na apresentação dos grafos.

5 Conclusão

Este trabalho teve como objetivo apresentar e demonstrar como a avaliação de consistência *DAG-dataset* pode ser realizada, sendo essa, fundamental para saber se o conjunto de dados gerado pelo processo causal está correto, ou seja, se o modelo causal estabelecido pelo DAG é consistente com o conjunto de dados observado.

Para isso, foram apresentados primeiramente, a terminologia dos DAGs, como esses são construídos, além das definições de independência condicional e a relação entre probabilidade e os grafos acíclicos dirigidos, estabelecendo assim, uma base para o entendimento da avaliação de consistência *DAG-dataset*.

Foram apresentadas duas ferramentas computacionais desenvolvidas por Textor et al. (2016), o *'DAGitty'* Web e o pacote `dagitty` disponível para uso no software R. Nessas ferramentas é possível desenhar o DAG, verificar o conjunto de ajustes suficientes e as implicações testáveis geradas pelo DAG, além disso, no pacote `dagitty` é possível realizar a avaliação de consistência *DAG-dataset*, sendo possível aplicar testes de independência condicional após a realização do critério de *d-separação*.

Uma breve introdução ao uso dessas ferramentas foi feita no Capítulo 4, apresentando suas principais funções. Com a aplicação de um exemplo de um estudo epidemiológico, apresentado em Kleinbaum et al. (1982), foi demonstrado o funcionamento do *'DAGitty'* Web e algumas funções disponíveis no pacote `dagitty`, incluindo como pode ser feita a avaliação de consistência *DAG-dataset* tanto para DAGs simples, quanto para DAGs de alta complexidade. O resultado obtido a partir deste exemplo foi que o DAG não é consistente com o seu *dataset*, possivelmente por possuir especificações incorretas das relações causais entre variáveis observadas presentes no DAG. Ainda, foi apresentada uma extensão do pacote `dagitty`, o pacote `ggdag`, de Barrett (2021), que pode ser usado em conjunto com o pacote `ggplot2`, sendo essa uma ótima opção para visualização gráfica do DAG e das relações causais do mesmo.

Estas ferramentas computacionais podem auxiliar, principalmente, pesquisadores que trabalham com DAGs com muitos vértices, facilitando a observação das implicações testáveis, assim como na avaliação de consistência do *DAG-dataset*. Al-

gumas das motivações principais para o uso destas ferramentas são sua praticidade, fácil entendimento e produção de resultados visualmente mais elegantes.

Em futuras contribuições para este assunto, a implementação de testes não-paramétricos quando não obtemos a suposição de normalidade das variáveis,ou seja, quando há uma associação não-linear entre as variáveis e possível multicolinearidade; visto que atualmente, o `dagitty` comporta apenas testes paramétricos e semi-paramétricos. E assim, comparar e avaliar o comportamento de testes de independência condicional paramétricos e não-paramétricos em diferentes cenários.

Referências Bibliográficas

Barrett, M. (2021). *ggdag: Analyze and Create Elegant Directed Acyclic Graphs*. R package version 0.2.3.

Cortes, T. R., Faerstein, E., e Struchiner, C. J. (2016). Utilização de diagramas causais em epidemiologia: um exemplo de aplicação em situação de confusão. *Caderno de Saúde Pública*, 32(8).

Greenland, S. e Pearl, J. (2007). Causal diagrams. *UCLA: Department of Statistics, UCLA. Disponível em:* https://escholarship.org/uc/item/7tn3p6jx.

Greenland, S., Pearl, J., e Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 10(1):37–48.

Kleinbaum, D. G., Kupper, L. L., e Morgenstern, H. (1982). *Epidemiologic research : principles and quantitative methods*. Belmont, CA: Lifetime Learning Publications.

Krieger, N., Sidney, S., e E, C. (1998). Racial discrimination and skin color in the cardia study: implications for public health research. coronary artery risk development in young adults. *Am J Public Health. doi:* 10.2105/ajph.88.9.1308. *PMID: 9736868; PMCID:* PMC1509091, Sep;88(9):1308–13.

Li, C. e Fan, X. (2019). On nonparametric conditional independence tests for continuous variables. *WIREs Computational Statistics*, page 12:e1489.

Pearl, J., Glymour, M., e Jewell, N. P. (2016). *Causal Inference in Statistics - A Primer*. Wiley.

Shipley, B. (2002). *Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations and Causal Inference*. Cambridge, UK: Cambridge University Press, 2002.

Tennant, P. W., Murray, E. J., Arnold, K. F., Berrie, L., Fox, M. P., Gadd, S. C., Harrison, W. J., Keeble, C., Ranker, L. R., Textor, J., Tomova, G. D., Gilthorpe,

M. S., e Ellison, G. T. (2020). Use of directed acyclic graphs (dags) to identify confounders in applied health research: review and recommendations. *International Journal of Epidemiology*, pages 1–13.

Textor, J., Zender, B. v. d., Gilthorpe, M. S., Liśkiewicz, M., e Ellison, G. T. (2016). Robust causal inference using directed acyclic graphs: the r package 'dagitty'. *International Journal of Epidemiology*, pages 1887–1894.

Wasserman, L. A. (2004). *All of statistics: a concise course in statistical inference*. Springer Texts in Statistics.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wright, S. (1920). The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs. *Proc Natl Acad Sci U S A. doi: 10.1073/pnas.6.6.320. PMID: 16576506; PMCID: PMC1084532*, 6(6):320–32.

Glossário

Confundimento É a distorção de uma medida do efeito de uma exposição em um desfecho devido a associação da exposição com outros fatores que influenciam na ocorrência do desfecho. O confundimento ocorre quando toda ou parte da associação aparente entre a exposição e o desfecho é de fato é determinada por outras variáveis que afetam o desfecho e não são afetadas pela exposição.

Desfecho Todos os resultados possíveis que podem decorrer da exposição a um fator causal ou de intervenções preventivas ou terapêuticas.

Exposição O fator que precede o desfecho, é condicionado a exposição com a causa ou possuir características de determinado problema de saúde. As definições de exposição podem incluir variáveis dicotômicas simples (por exemplo, sempre exposto vs. nunca exposto) ou ser mais detalhadas, incluindo estimativas de duração, janelas de exposição (por exemplo, exposição atual vs. passada) ou dosagem (por exemplo, dosagem atual, dosagem cumulativa ao longo do tempo). Também chamado de fator de estudo, variável preditora ou variável independente.

Viés de colisão Associação não causal entre as causas diretas compartilhadas do colisor (ou seja, seus ancestrais), quando o ajuste (ou condicionamento) é feito em um colisor.

Viés de confusão Viés do efeito estimado de uma exposição em um desfecho devido à presença de causas comuns da exposição e do desfecho.