Predição e Causalidade: utilizando aprendizado de máquina na inferência causal*

Rodrigo Moita†

Ricardo Ceneviva‡

Resumo

A pesquisa em ciências sociais abrange quatro tarefas analíticas inter-relacionadas: descrição, compreensão substantiva, explicação causal e previsão. Muitos estudos não distinguem claramente essas tarefas, o que pode levar a um desalinhamento entre referenciais teóricos, desenho da pesquisa e métodos estatísticos. Neste relatório, argumenta-se que definir precisamente o estimando e diferenciar essas tarefas é crucial para inferências confiáveis. Revisam-se três abordagens clássicas de inferência causal – variáveis instrumentais, descontinuidade de regressão e diferenças em diferenças – ilustrando os pressupostos e o estimando alvo de cada método com o exemplo da relação entre educação e renda. Examina-se também como técnicas de aprendizado de máquina podem complementar a análise causal, melhorando a estimação e revelando heterogeneidades de efeito de tratamento, salientando que essas técnicas não substituem um delineamento de pesquisa rigoroso nem fortes suposições de identificação. Ao enfatizar a comunicação transparente de suposições, diagnósticos minuciosos e clareza sobre o escopo inferencial, destacam-se as melhores práticas para aprimorar a reprodutibilidade e a validade dos resultados. Por fim, argumenta-se que integrar uma descrição acurada, profundidade teórica, metodologia causal rigorosa e modelagem preditiva gera um conhecimento mais cumulativo e socialmente relevante, especialmente na era do big data e da ciência social computacional.

Introdução

Questões de pesquisa em Ciências Sociais podem ser agrupadas em quatro tarefas analíticas inter-relacionadas: descrição, compreensão substantiva, explicação causal e predição. A descrição procura responder "o quê" e "quanto", capturando regularidades empíricas e quantificando fenômenos sociais de forma parsimoniosa. A compreensão substantiva busca interpretar e atribuir significados a essas regularidades, propondo mecanismos teóricos subjacentes (p. ex., hipóteses de por que certos padrões ocorrem). A explicação causal visa determinar relações de causa e efeito, respondendo a perguntas "o quê causa o quê" por meio de comparações contrafactuais sob suposições explícitas e diagnosticáveis. Já a predição foca em prever valores de um desfecho para novas observações, avaliando o desempenho fora da amostra, sem necessariamente esclarecer mecanismos causais. Essas tarefas são complementares, porém distintas, e reconhecê-las é crucial para progresso científico (Shmueli, 2010; Breiman, 2001). Modelos estatísticos com alto poder explicativo (causal) nem sempre têm alto poder preditivo, e vice-versa (Shmueli, 2010).

Nas ciências sociais, tradicionalmente predominou o uso de modelos para explicação causal, ao passo que a predição foi relegada a segundo plano, gerando confusão entre os objetivos

*Trabalho apresentado no 53º Encontro Nacional de Economia, São Paulo/SP, INSPER, 16 a 19/12/2025.

†Professor titular do Departamento de Economia da Faculdade de Economia, Administração e Contabilidade da (FEA) da Universidade de São Paulo (USP)

‡Professor visitante do Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas (CECS) da Universidade Federal do ABC (UFABC)

explicativos e preditivos (Shmueli, 2010; Breiman, 2001). Essa distinção, contudo, "deve ser entendida para o avanço do conhecimento científico" (Shmueli, 2010). Além disso, a descrição muitas vezes é desdenhada como "mera estatística descritiva", apesar de ser base para teorias e hipóteses causais (Grimmer, 2015). De fato, medir cuidadosamente fenômenos sociais – quem faz o quê, quanto e quando – fornece os fatos estilizados que motivam modelos teóricos e perguntas causais (Grimmer, 2015). Ignorar a diferença entre descrever, explicar causalmente, compreender mecanismos e predizer pode levar a inferências equivocadas e uso inadequado de métodos (Shmueli, 2010; Morgan; Winship, 2015).

Estimandos, teoria e desenho de pesquisa

Cada tarefa analítica requer definir com precisão o estimando de interesse – isto é, a quantidade-alvo a ser estimada – e alinhar o desenho de pesquisa e os critérios de validação a esse estimando (Lundberg; Johnson; Stewart, 2021). Em outras palavras, antes de coletar dados ou ajustar modelos, o pesquisador deve responder claramente: "Qual é a quantidade (parâmetro populacional) que pretendo estimar?" (Lundberg; Johnson; Stewart, 2021). O estimando conecta a evidência estatística à teoria substantiva (Morgan; Winship, 2015; Pearl, 2010). Por exemplo, pode ser um contraste descritivo (como a diferença média de salários entre grupos), um parâmetro causal (como o efeito médio de educação sobre salário) ou um critério preditivo (como o erro quadrático médio de um modelo de previsão). Definir o estimando fora de qualquer modelo estatístico específico evita que a questão de pesquisa fique tacitamente limitada a parâmetros de um modelo particular (como coeficientes de uma regressão linear) (Lundberg; Johnson; Stewart, 2021; Angrist; Pischke, 2009). Pelo contrário, explicitar "a quantidade central de cada análise – o estimando teórico – em termos precisos que existem fora do modelo" permite escolher métodos apropriados e comunicar claramente o objetivo da inferência (Morgan; Winship, 2015; Pearl, 2010). Muitas vezes, cientistas sociais "pulam a etapa de definir o estimando" e saltam direto aos dados e métodos; sem o estimando declarado, fica impossível ao leitor julgar se os procedimentos e suposições adotados são adequados (Lundberg; Johnson; Stewart, 2021; Cunningham, 2021). Em suma, um delineamento de pesquisa rigoroso passa por: (1) estabelecer o estimando teórico, conectado à teoria e indicando claramente o mecanismo ou contrafactual de interesse e a população-alvo; (2) vincular esse estimando teórico a um estimando empírico que possa ser obtido dos dados sob certas suposições de identificação; e (3) escolher um método (estimador) e coletar dados para obter a estimativa desejada (Morgan; Winship, 2015; Lundberg; Johnson; Stewart, 2021).

Nota metodológica Convém distinguir os termos estimando, estimador e estimativa. O estimando é a quantidade populacional alvo (por exemplo, a diferença média de salários se todos tivessem ensino superior versus se todos tivessem apenas ensino médio). O estimador é o procedimento ou fórmula estatística usada para estimar essa quantidade (por exemplo, a diferença de médias amostrais ou um coeficiente de regressão). Já a estimativa é o valor numérico obtido na amostra (por exemplo, um aumento estimado de 20% no salário médio). Tornar explícito o estimando e separá-lo do estimador previne que se confunda a pergunta substantiva com a técnica utilizada (Lundberg; Johnson; Stewart, 2021; Morgan; Winship, 2015). Essa separação também reforça a clareza na comunicação: ajuda leitores a entender quais suposições ligam a evidência ao estimando teórico e quais escolhas metodológicas foram feitas para obtê-lo (Lundberg; Johnson; Stewart, 2021; Morgan; Winship, 2015).

Descrição e compreensão substantiva

O primeiro passo em muitas pesquisas é descritivo: quantificar e resumir os dados relevantes, estabelecendo fatos estilizados. Por exemplo, consideremos o fio condutor desta discussão, a relação entre educação e salários. Descritivamente, podemos perguntar: qual é o prêmio salarial médio associado a maior escolaridade? Uma análise descritiva típica é a regressão linear de mínimos quadrados ordinários (MQO/OLS) de log-salário contra anos de educação e outras covariáveis básicas. Essa regressão frequentemente encontra que cada ano adicional de estudo está associado a um aumento de, digamos, 6–10% no salário médio (Angrist; Pischke, 2009). Esse valor – conhecido como retorno educacional observado – descreve "o quê" está presente nos dados: em média, pessoas mais educadas ganham salários mais altos. Entretanto, a interpretação substantiva dessa regularidade requer compreensão teórica: por que educação e renda se correlacionam? Diferentes mecanismos foram propostos na literatura. A teoria do capital humano sugere que a educação aumenta a produtividade do trabalhador (por meio de habilidades e conhecimento), levando a salários maiores. A teoria da sinalização, por sua vez, argumenta que diplomas servem como credenciais que sinalizam habilidade ou diligência aos empregadores, que então pagam mais aos diplomados. Outros mecanismos podem incluir ampliação de redes sociais, acesso a melhores oportunidades ou exigências de credenciais no mercado de trabalho. Essa compreensão substantiva contextualiza a descrição: embora os dados mostrem correlação educação–salário, precisamos da teoria para explicar o como e por que dessa relação – isto é, quais processos sociais geram o padrão observado.

É importante notar que descrição e compreensão não estabelecem por si causalidade. A correlação positiva entre escolaridade e rendimento pode refletir, além dos mecanismos acima, fatores de confusão. Por exemplo, alunos de maior habilidade ou origem socioeconômica privilegiada tendem a obter mais anos de estudo e maiores salários independentemente da educação. Assim, parte do "prêmio salarial" descritivo pode advir de diferenças preexistentes entre indivíduos, e não do efeito da educação em si. Em nosso exemplo, se estudantes de habilidade cognitiva elevada permanecem mais tempo na escola, a diferença salarial observada de 6–10% ao ano de estudo inflaria o verdadeiro efeito causal da educação, pois incluiria o "prêmio" da maior habilidade (variável omitida) (Morgan; Winship, 2015). Esse fenômeno é o viés de variável omitida: a estimativa descritiva não é informativa sobre o efeito de uma política de elevar a escolaridade se características como habilidade não forem devidamente consideradas (Cunningham, 2021). Portanto, a transição da descrição para a explicação causal exige cuidado: precisamos isolar, tanto quanto possível, o impacto da educação sobre salários de possíveis fatores de confusão.

Explicação causal: conceitos e inferência clássica

Explicar causalmente implica responder a perguntas do tipo "qual seria o salário de um indivíduo se sua escolaridade fosse diferente do que de fato é?". Formalmente, define-se $Y(0)$ como o resultado (salário, por exemplo) que um indivíduo teria sob uma condição de referência (como educação baixa) e $Y(1)$ sob uma condição alternativa (educação alta). O efeito causal individual da educação é a diferença $Y(1) - Y(0)$. Como nunca observamos ambos os cenários para o mesmo indivíduo, lidamos com o problema fundamental da inferência causal – a falta do contrafactual para cada unidade. A solução envolve comparação entre grupos ou momentos distintos, sob suposições que permitam interpretar essa comparação como se fosse o contrafactual. No modelo de resultados potenciais (Neyman–Rubin), o efeito causal médio é definido pelo ATE (Average Treatment Effect) – por exemplo, o efeito médio de ter ensino superior versus não tê-lo sobre o salário na população alvo. Podemos também definir o ATT (Average Treatment effect on the Treated) – o efeito mé-

dio do tratamento para aqueles que efetivamente receberam educação superior – e o ATU (efeito médio para os não tratados); outro parâmetro importante é o LATE (Local Average Treatment Effect), o efeito médio em um subgrupo específico definido por um instrumento ou corte (discutido adiante). Em linguagem simples, o ATE responde "em média, qual o impacto causal do aumento de escolaridade sobre o salário?", enquanto o LATE poderia responder "qual é o efeito causal para indivíduos cuja escolaridade foi aumentada devido a uma certa intervenção (p.ex., uma mudança na lei)?".

Duas suposições centrais para identificar causalidade em dados observacionais são a ignorabilidade (ou não-confundimento) e a sobreposição (suporte comum). Ignorabilidade significa que, condicional em um conjunto de covariáveis observadas X (como habilidade, origem familiar, etc.), a atribuição do "tratamento" (educação alta vs. baixa) é "como se aleatória", ou seja, independente dos resultados potenciais $Y(0), Y(1)$. Em outras palavras, após controlar por X, não há diferenças sistemáticas entre os grupos que afetem os resultados além do tratamento. Já a sobreposição requer que para cada combinação de covariáveis X haja presença tanto de indivíduos tratados quanto de não tratados – garantindo comparabilidade; não pode haver, por exemplo, um grupo que sempre recebe educação alta independentemente de X. Sob essas condições, pode-se estimar o efeito causal médio ajustando-se para X. Em termos de gráficos causais de Pearl, ignorabilidade equivale a dizer que não há caminhos de *backdoor* não bloqueados entre a causa e o efeito após ajustar X. Pearl formalizou o critério de *backdoor*, estipulando que é suficiente controlar um conjunto S de covariáveis que (i) não sejam consequências da causa X e (ii) bloqueiem todos os caminhos causais alternativos de X até Y (Pearl, 2010). Se tais S existirem e forem medidos, o efeito de X sobre Y fica identificável a partir de dados observacionais (Pearl, 2010). Em suma, se conseguirmos "tornar comparáveis" os indivíduos nos cenários alternativos através do design (seja randomizando ou controlando fatores de confusão medidos), podemos inferir causação.

Na prática, atingir ignorabilidade estrita é desafiador – quase sempre há algum fator não observado influenciando tanto a educação quanto o salário (ex.: motivação, redes de contato). Por isso, métodos de inferência causal foram desenvolvidos para aproximar as condições de um experimento. Três delineamentos clássicos em econometria e ciências sociais para identificar efeitos causais com dados observacionais são: Variáveis Instrumentais (VI), Descontinuidade de Regressão (RD) e Diferenças em Diferenças (DiD). Cada estratégia almeja um estimando causal sob um conjunto distinto de suposições, frequentemente mais plausível que a ignorabilidade "pura" em certos contextos. A seguir, apresentamos cada estratégia aplicando ao exemplo educação–salário, ressaltando suposições identificadoras, estimandos locais produzidos e diagnósticos mínimos de validade.

Variáveis Instrumentais (VI). Suponha que temos um fator Z (instrumento) que afeta a educação de uma pessoa, mas não afeta diretamente seu salário, exceto por via da educação. Por exemplo, uma mudança na lei de escolaridade obrigatória que eleva a idade mínima de abandono escolar de 14 para 15 anos. Tal mudança (implementada no Reino Unido em 1947) forçou alunos nascidos após certa data a permanecer mais um ano na escola. Nesse caso, $Z =$ "nascer após a data de corte (coorte afetada pela lei)" e $X =$ anos de educação. A suposição é que Z impacta Y (salário adulto) apenas por alterar X (educação), e não por outros caminhos – esta é a restrição de exclusão. Além disso, Z deve estar correlacionado com X (a relevância do instrumento), ou seja, de fato a lei aumentou os anos de estudo da coorte afetada. No exemplo, a evidência mostra que a reforma elevou a escolaridade média e produziu um aumento nos salários médios da coorte afetada; a razão entre essas variações – estimativa de Wald – sugere um efeito causal local positivo por ano adicional de estudo induzido pela lei (Angrist; Krueger, 1991; Imbens; Angrist, 1994; Angrist; Pischke, 2009). Esse é um LATE: o efeito médio causal da educação para

aqueles indivíduos que só estudaram mais por causa da mudança de lei, isto é, os chamados *compliers*. Importante notar: (i) encontrar instrumentos válidos é difícil; é preciso argumentação teórica e evidências de que Z não afeta Y por caminhos alternativos e que não haja "violadores" significativos; (ii) o efeito estimado via VI geralmente não vale para toda a população; vale para o subgrupo afetado pelo instrumento (Imbens; Angrist, 1994; Angrist; Pischke, 2009). Diagnósticos mínimos incluem verificar a força do instrumento (estatística-F do primeiro estágio) e discutir plausibilidade da exclusão.

Descontinuidade de Regressão (RD). Estruturas de RD exploram situações em que uma regra administrativa cria um corte em uma variável contínua, atribuindo tratamento a quem fica de um lado do limiar e não ao outro. Próximo ao ponto de corte, assume-se que unidades são comparáveis, exceto pelo tratamento. A mudança legal de idade mínima pode ser vista como um desenho RD em torno da data de corte: comparar indivíduos imediatamente de um lado e outro fornece uma estimativa causal local do efeito de um ano extra de educação. A suposição-chave é a continuidade: na ausência do tratamento, não haveria salto no desfecho exatamente no ponto de corte. Evidências gráficas e testes estatísticos ajudam a sustentar a suposição (Angrist; Pischke, 2009). O estimando aqui é o efeito médio local no ponto de corte.

Diferenças em Diferenças (DiD). Quando uma intervenção afeta apenas parte de uma população em um determinado tempo, comparamos a evolução do desfecho entre o grupo tratado e um grupo controle, assumindo tendências paralelas na ausência de tratamento. O estimando típico é o ATT no período pós-intervenção. Validações incluem checagens de tendências prévias e placebos (Cunningham, 2021; Angrist; Pischke, 2009).

Em todos os delineamentos, é fundamental explicitar o estimando e o escopo de validade. No OLS descritivo, a associação média pode estar viesada por confusão. Em VI e RD, o estimando é local (LATE/efeito no limiar); em DiD, é o ATT para o grupo tratado. Esses efeitos podem diferir do ATE populacional. Resultados devem explicitar "qual efeito foi estimado e para quem" (Angrist; Pischke, 2009; Morgan; Winship, 2015).

Definição: Causalidade

O raciocínio causal moderno se ancora em duas formulações:

- O modelo de resultados potenciais (Rubin; Holland);
- A modelagem gráfica de Pearl, baseada em DAGs.

O ponto de partida é o problema fundamental da inferência causal: nunca observamos o mesmo indivíduo nos dois estados possíveis (tratado e não tratado). Por isso, a causalidade deve ser inferida por meio de comparações entre unidades observáveis que satisfaçam ignorabilidade e sobreposição. A suposição de ignorabilidade condicional estabelece que, dado um conjunto de covariáveis observadas Z que bloqueia todos os caminhos de confusão, os resultados potenciais são independentes da atribuição ao tratamento. Sob forte ignorabilidade, requer-se também positividade e SUTVA. Em experimentos, a aleatorização garante a ignorabilidade; em estudos observacionais, sua plausibilidade depende da adequação de Z e da qualidade de mensuração. Diagnósticos empíricos – balanceamento de covariáveis, verificação de sobreposição – e análises de sensibilidade ajudam a avaliar robustez. Quando tais condições são frágeis, recorrem-se a desenhos alternativos (VI, RD, DiD) (Morgan; Winship, 2015; Pearl, 2010; Cunningham, 2021).

Predição e modelos de aprendizado de máquina (ML) nas ciências sociais

Diferente da explicação causal, cujo foco é estimar parâmetros interpretáveis ligados a mecanismos, a predição visa maximizar a capacidade de prever Y dado X, avaliada por métricas de erro preditivo fora da amostra. Leo Breiman distinguiu duas "culturas" na modelagem estatística: a cultura orientada a modelos estocásticos e a cultura algorítmica (Breiman, 2001). Na cultura tradicional, assume-se um modelo paramétrico e foca-se em interpretação; na cultura algorítmica, busca-se acurácia preditiva sem exigir interpretabilidade. Essa diferença reflete objetivos distintos: explicar versus prever (Shmueli, 2010).

Consequentemente, os critérios de sucesso divergem: em explicação, avaliam-se suposições e significância; em predição, o critério é o desempenho em novos dados. Um modelo causal bem especificado pode não ser o mais preditivo; e um modelo preditivo ótimo pode não isolar mecanismos causais (Shmueli, 2010). Na era dos "big data", essa distinção ganhou relevo: áreas de ciência de dados desenvolveram modelos altamente preditivos, enquanto as ciências sociais enfatizaram estratégias causais robustas. Hoje, reconhece-se complementaridade entre ambas (Grimmer, 2015; Brand; Zhou; Xie, 2023). Registros administrativos massivos e rastros digitais permitem novas medições e heterogeneidades, mas não resolvem a identificação causal sem desenho apropriado (Grimmer, 2015; Salganik, 2019).

Uma manifestação concreta dessa convergência é o surgimento de métodos de ML causal. Estimadores duplamente robustos e *Double/Debiased Machine Learning* permitem usar algoritmos flexíveis para estimar componentes de viés e, via *cross-fitting*, obter efeitos com boas propriedades assintóticas (Chernozhukov *et al.*, 2018; Brand; Zhou; Xie, 2023). Métodos de heterogeneidade como árvores/ florestas causais estimam CATEs e revelam variação de efeitos (Athey; Imbens, 2016). Nada disso elimina a necessidade de suposições identificadoras: ignorabilidade, instrumentos válidos ou variação quase-experimental continuam centrais (Morgan; Winship, 2015; Pearl, 2010).

Retomando educação → salários sob a ótica preditiva e ML-causal: um modelo de ML pode prever salários com alta acurácia usando muitas variáveis, mas isso não implica causalidade. Abordagens duplamente robustas podem isolar o efeito da educação sob suposições explícitas. Explorar heterogeneidades pode orientar políticas, mas requer cautela com confundimento em subgrupos e múltiplas comparações (Brand; Zhou; Xie, 2023; Cunningham, 2021).

Comunicação científica e boas práticas de transparência

Independentemente da abordagem, pesquisas sólidas requerem clareza sobre suposições, escopo inferencial, limitações e reprodutibilidade. Idealmente, cada estudo deveria "declarar seu estimando" explicitamente (Lundberg; Johnson; Stewart, 2021). Em estudos causais observacionais, deve-se ser transparente sobre as suposições de identificação e reportar diagnósticos que aumentem a confiança nelas – por exemplo, testes de equilíbrio, estatística-F do primeiro estágio, gráficos de RD e verificação de tendências paralelas (Angrist; Pischke, 2009; Cunningham, 2021). É crucial delimitar o escopo: estimativas locais podem não generalizar para outros contextos (Brady, 2008).

Práticas de reprodutibilidade incluem disponibilizar, quando possível, dados e códigos, *pré-registro* e material suplementar (Salganik, 2019). Em ciência social computacional, isso é especialmente relevante devido à complexidade dos dados e dos procedimentos de ML (Grim-

mer, 2015). Comunicar limites, evitar inferência causal indevida em análises descritivas e evitar extrapolações não sustentadas fortalece a confiabilidade do trabalho. Integrando descrição, teoria, inferência causal, predição e transparência, avançamos em uma ciência social cumulativa e socialmente relevante (Chattopadhyay; Zubizarreta, 2024).

Modelos Lineares e Modelos Lineares Generalizados

A regressão linear ocupa um lugar central no campo da análise multivariada e é reconhecida como uma das técnicas estatísticas mais difundidas nas ciências sociais (SlideShare, s.d.). Seu apelo reside na capacidade de modelar relações entre múltiplas variáveis de forma relativamente simples e interpretável. Em essência, a regressão linear busca quantificar como mudanças em uma ou mais variáveis explicativas se associam a variações em uma variável resultado, mantendo constantes os demais fatores no modelo. Por exemplo, um pesquisador pode investigar em que medida anos de escolaridade influenciam o salário de um indivíduo. Essa relação entre educação e rendimentos é um fio condutor clássico na economia do trabalho e será o exemplo integrador deste relatório. Antes de aplicar qualquer modelo, é fundamental definir com clareza o que se quer estimar – isto é, qual é o efeito ou parâmetro populacional de interesse (o estimando) (Lundberg; Johnson; Stewart, 2021). No nosso caso, o estimando pode ser definido como o efeito médio de uma ano adicional de escolaridade sobre o salário (em termos percentuais). Convém distinguir esse conceito de estimador e estimativa: o estimador é a técnica ou fórmula empregada (aqui, o método dos mínimos quadrados ordinários – MQO) e a estimativa é o valor numérico obtido na amostra. Segundo Lundberg, Johnson e Stewart (2021), "o estimando é a quantidade-alvo – o propósito da análise estatística" (Lundberg; Johnson; Stewart, 2021), devendo ser estabelecido em termos substantivos e teóricos antes mesmo da estimação. A regressão linear simples ou múltipla, então, é uma ferramenta para obter um estimador para esse efeito, sob certas suposições. A popularidade da regressão linear nas ciências sociais também se justifica por sua extensibilidade e diagnóstico: o modelo linear serve de base para métodos mais complexos (como modelos lineares generalizados, regressões multinível, séries temporais, etc.) e vem acompanhado de um arcabouço robusto de técnicas de diagnóstico de adequação do modelo. Além disso, os coeficientes da regressão têm uma interpretação direta (como veremos, indicando variações médias na variável dependente associadas a variações unitárias nas explicativas), o que facilita a comunicação dos resultados a públicos não técnicos. Em suma, a regressão linear é frequentemente a "porta de entrada" para análise multivariada por combinar rigor matemático com relativa simplicidade conceitual, sendo "indiscutivelmente a técnica estatística mais utilizada nas ciências sociais" (SlideShare, s.d.). Sintese: a regressão linear não é apenas uma técnica; é uma linguagem comum da inferência empírica nas ciências sociais. Neste relatório, organizamos a exposição de forma didática seguindo a sequência lógica de um estudo empírico: apresentação do modelo e do exemplo substantivo (educação e salário), desenvolvimento da intuição do modelo linear simples e múltiplo, explicitação dos pressupostos e métodos de diagnóstico, interpretação cuidadosa dos coeficientes e da qualidade do ajuste, discussão dos limites (especialmente no tocante à causalidade e endogeneidade) e, por fim, implicações para comunicação de resultados e boas práticas de pesquisa.

Modelo de Regressão Linear Simples: educação → log do salário

Para iniciar, consideremos um modelo de regressão linear simples examinando a relação entre escolaridade e salário. Traduzimos o salário para escala logarítmica (log-salário)

por dois motivos: (1) a distribuição de salários costuma ser assimétrica à direita (alguns indivíduos ganham muito mais que a maioria), e o logaritmo tende a aproximar a distribuição da normalidade e estabilizar a variância; (2) a interpretação dos coeficientes torna-se percentual – coeficientes em um modelo com logaritmo do salário como dependente indicam aproximadamente a variação percentual no salário associada a uma mudança de uma unidade na variável explicativa. Formalmente, especificamos:

$$\log(\text{salário}_i) = \beta_0 + \beta_1 \cdot \text{educação}_i + u_i,$$

onde educação_i representa os anos de escolaridade do indivíduo i, $\log(\text{salário}_i)$ é o logaritmo (natural) de seu salário, β_0 é o intercepto (valor esperado de log-salário para alguém sem escolaridade, em teoria) e β_1 é o coeficiente de interesse, associado à variável educação. O termo u_i (às vezes denotado ε_i) é o termo de erro aleatório, que agrega todos os fatores não incluídos explicitamente no modelo e quaisquer flutuações individuais. Esse modelo assume uma forma funcional linear, ou seja, que a relação entre educação e o log do salário pode ser bem aproximada por uma reta. Intuitivamente, β_1 mede a diferença média no log-salário para cada ano adicional de estudo. Se, por exemplo, $\beta_1 = 0{,}08$, isso sugere que um ano a mais de educação está associado, em média, a um salário cerca de 8% mais alto ($e^{0{,}08} \approx 1{,}083$; aproximadamente 8,3% de aumento). Trata-se de uma interpretação substantiva valiosa: estamos quantificando o retorno econômico da educação em termos percentuais médios. Antes de estimar β_1 com os dados, um pesquisador prudente realiza análises descritivas e inspeção gráfica. Estatísticas como média, mediana e desvio-padrão de anos de escolaridade e salários fornecem um panorama da amostra. É útil conhecer, por exemplo, qual o salário médio (ou mediano) e como varia conforme níveis de educação. Uma visualização inicial típica seria um diagrama de dispersão do salário (ou log-salário) em função dos anos de estudo, possivelmente com uma linha de tendência ajustada. Essa inspeção pode revelar se a relação parece aproximadamente linear, se há outliers (por exemplo, alguém com baixa educação mas salário muito alto, o que pode indicar um caso peculiar), ou se há indícios de heterogeneidade que demandariam transformações (por exemplo, salário crescendo de forma exponencial mais do que linear). Na narrativa que nos guia, imaginemos que dispomos de dados de uma pesquisa amostral de adultos em idade ativa. A inspeção gráfica mostra uma correlação positiva: em geral, indivíduos com mais anos de estudo tendem a apresentar salários logarítmicos maiores. Em outras palavras, educação e salário exibem associação positiva, consistente com a teoria do capital humano. Essa observação inicial prepara terreno para a regressão: esperamos um β_1 positivo. Contudo, a regressão linear simples é apenas um ponto de partida. O termo de erro u_i captura tudo mais que influencia salário além de educação – por exemplo, experiência profissional, habilidades pessoais, gênero, região geográfica, setor de emprego, entre outros. Se alguma dessas características omitidas estiver correlacionada com educação, a estimação de β_1 pode ser enviesada. Por isso, avançamos para o modelo de regressão múltipla, introduzindo variáveis de controle que representem fatores importantes.

Regressão Múltipla e o Papel do Controle

A regressão linear múltipla estende o modelo para incluir diversas variáveis explicativas. No contexto do nosso exemplo, podemos adicionar controles clássicos de equações de rendimentos, tais como experiência no mercado de trabalho (anos de experiência), gênero (por exemplo, uma variável indicadora para feminino/masculino) e região (dummies para grandes regiões ou área urbana/rural). O modelo poderia ser escrito como:

$$\log(\text{salário}_i) = \beta_0 + \beta_1 \cdot \text{educação}_i + \beta_2 \cdot \text{experiência}_i + \beta_3 \cdot \text{gênero}_i + \beta_4 \cdot \text{região}_i + u_i.$$

Cada coeficiente β_j agora representa um efeito condicional: β_1 é a diferença média no log-salário por ano adicional de educação mantendo constante a experiência, o gênero e a região do indivíduo. Em outras palavras, a regressão múltipla permite isolar a associação entre educação e salário dos efeitos de outros fatores, aproximando a ideia de "comparar o comparável". Ao controlar a variável gênero, por exemplo, comparamos homens e mulheres com níveis equivalentes de educação e experiência; ao controlar a experiência, comparamos trabalhadores com escolaridade diferente mas tempo similar de mercado, e assim por diante. Na prática, a introdução de controles muitas vezes provoca uma mudança na magnitude (e eventualmente no significado) do coeficiente de educação. Se β_1 era, digamos, 0,10 (10% de retorno por ano) no modelo simples, ele pode se reduzir para 0,07 (7%) ao incluir controles, caso parte da associação inicial se devesse a fatores correlacionados. Por exemplo, pessoas com mais escolaridade talvez sejam mais jovens e tenham menos experiência (em certos contextos históricos), o que mascara um pouco seu salário; ao controlar experiência, o efeito "puro" da educação pode ficar mais evidente (aumentando β_1) ou, alternativamente, se indivíduos mais escolarizados tendem a trabalhar em regiões metropolitanas com maiores salários, e a variável região capta esse efeito, o coeficiente de educação pode diminuir, refletindo o desconto daquele componente que na verdade era devido à localização. Esse exercício de adicionar variáveis ao modelo nos alerta sobre a importância de pensar no desenho da pesquisa e no estimando alvo. Ao incluir um conjunto de controles, estamos efetivamente alterando a pergunta que o modelo responde. O coeficiente β_1 no modelo múltiplo responde: "qual é a diferença média no salário entre duas pessoas que diferem em um ano de estudo, mas são similares em experiência, gênero e região?". Essa é uma estimativa de associação condicional – um passo mais próximo de causalidade do que a associação bruta, mas ainda dependendo de suposições fortes (como discutiremos adiante). Entretanto, nem todos os controles são benéficos; incluir variáveis irrelevantes aumenta a variância das estimativas sem reduzir viés, e incluir controles que sejam consequências da variável principal ou altamente correlacionados a ela pode introduzir complicações (os chamados "bad controls" na literatura econométrica) (UCLA, s.d.). No nosso exemplo, experiência e região são plausíveis confundidores a controlar, mas controle por salário do pai, por exemplo, embora correlacionado com educação, seria problemático pois é um proxy de origem socioeconômica e possivelmente já incorpora parte do efeito que a educação teria (além de poder estar no caminho causal, dependendo do argumento). Em suma, a escolha de controles deve ser guiada por teoria e lógica causal. Com o modelo múltiplo estimado, supomos que encontramos $\beta_1 \approx 0{,}08$ (8% por ano, por exemplo). Essa magnitude pode agora ser interpretada com mais confiança como o retorno médio da educação, condicionado a outros fatores observáveis. Ainda assim, precisamos verificar se o modelo ajustado atende aos pressupostos do método de regressão e se os resultados são estatisticamente robustos. Entramos, assim, no terreno dos pressupostos e diagnóstico do modelo linear.

Pressupostos do Modelo Linear e Diagnóstico

Para que as inferências derivadas de um modelo de regressão linear sejam válidas – especialmente ao usar MQO (mínimos quadrados ordinários) – é necessário que certos pressupostos (suposições) sejam aproximadamente satisfeitos (Stats Made Easy, s.d.). Estes pressupostos dizem respeito tanto à forma funcional quanto às propriedades dos erros do modelo e dos dados. Podemos elencar os principais: **Linearidade:** Supõe-se que a relação entre cada variável explicativa e a variável dependente seja linear aditiva. Isto é, a contribuição de cada X em Y é linear e soma-se à dos demais (sem termos multiplicativos ou exponenciais, a menos que explicitamente incluídos como variáveis no modelo). É possível verificar a linearidade inspecionando gráficos de dispersão e, após a estimação, gráficos de resíduos versus valores ajustados. Se houver padrões não-lineares nos resíduos (por exem-

plo, curvatura), isso indica violação da linearidade (Stats Made Easy, s.d.) – nesse caso, transformações (como incluir um termo quadrático de educação, ou usar log em alguma variável) podem ser necessárias. **Independência das observações (ausência de autocorrelação):** Assume-se que cada observação é independente das demais. Em amostras aleatórias simples de pesquisas sociais, isso geralmente é plausível. No entanto, se os dados tiverem estrutura de dependência (por exemplo, alunos dentro da mesma escola, famílias dentro do mesmo bairro, ou dados coletados ao longo do tempo do mesmo indivíduo), as suposições de independência e iid (identicamente distribuídos) são violadas. A independência é particularmente relevante quanto aos erros: pressupomos que não haja correlação entre u_i e u_j (nenhum padrão sistemático não-modelado ligando observações). Quando há autocorrelação (por exemplo, em séries temporais econômicas), técnicas específicas ou ajustes nos erros são necessários. Neste relatório focamos no caso cross-section simples, pressupondo independência entre indivíduos da amostra. **Homoscedasticidade (variância constante dos erros):** Um pressuposto fundamental da regressão linear múltipla é que a variância dos resíduos permaneça constante ao longo de todas as observações (Stats Made Easy, s.d.). Em termos simples, os erros u_i devem ter aproximadamente a mesma variância independentemente dos valores de X. Se em faixas diferentes de escolaridade o erro (isto é, a variabilidade salarial não explicada pelo modelo) for muito diferente – por exemplo, salários de pessoas com baixa educação variam pouco entre si, ao passo que entre pessoas com alta educação variam muito – há heterocedasticidade. Podemos diagnosticar heterocedasticidade inspecionando um gráfico de resíduos versus valores ajustados: se os resíduos formam um "funil" (espalhando-se mais em um extremo do que no outro), é indício de variância não constante (Stats Made Easy, s.d.). Testes formais como o teste de Breusch-Pagan ou White também podem detectar essa condição. A homocedasticidade é importante porque, quando satisfeita (juntamente com outras condições), os estimadores MQO são os mais eficientes entre os estimadores lineares não-viesados (teorema de Gauss-Markov) (Stats Made Easy, s.d.). Se violada, os coeficientes estimados ainda são insesgados (desde que outros pressupostos se mantenham), porém as estatísticas de inferência (erro padrão, t e p-valores) podem estar incorretas. Remédios: Uma vez detectada a heterocedasticidade, podemos usar erros-padrão robustos (ajustados) que corrigem a inferência sem mudar os coeficientes; também é possível reespecificar o modelo (por exemplo, verificar se falta alguma variável cuja omissão cause variância do erro diferente) ou aplicar pesos aos casos (regressão por Mínimos Quadrados Ponderados) caso identifiquemos a fonte da heterogeneidade da variância (Stats Made Easy, s.d.). Transformar variáveis (como usar log do salário, o que já fizemos) é uma estratégia comum que frequentemente diminui a heterocedasticidade. **Normalidade dos erros:** Assume-se que, condicional às variáveis explicativas, os erros u_i sigam aproximadamente uma distribuição normal. Esse pressuposto, na verdade, é necessário principalmente para validação de inferências estatísticas (intervalos de confiança e testes de hipóteses), pois garante que os estimadores tenham distribuição exata conhecida em pequenas amostras. Em amostras grandes, pelo Teorema Central do Limite, a média amostral de erros tende à normalidade mesmo que os erros individuais não sejam normais, de modo que esse pressuposto se torna menos crítico (os estimadores MQO serão assintoticamente normais sob condições gerais). No entanto, verificar a normalidade dos resíduos – por meio de um histograma ou Q-Q plot dos resíduos – é boa prática. Uma forte assimetria ou caudas muito pesadas podem sugerir que alguma observação discrepante ou variável omitida esteja afetando o ajuste. Se os resíduos não forem nem aproximadamente normais, podemos recorrer a transformações (como já mencionado) ou, se a não-normalidade advém de outliers, considerar técnicas robustas (embora para fins didáticos não entraremos nelas aqui). Importa frisar: leve violação da normalidade raramente invalida um modelo; a preocupação maior é se a não-normalidade é sintoma de mau ajuste (p. ex., omissão de variável importante levando a resíduos estruturados).

Não multicolinearidade perfeita: No contexto de regressão múltipla, requer-se que nenhuma variável explicativa seja combinação linear exata de outra(s) – em termos práticos, não pode haver correlação perfeita entre duas variáveis regressoras. Se, por exemplo, "anos de estudo" e "grau de instrução" (em categorias) estivessem juntos no modelo, haveria uma relação determinística entre eles. Nesses casos de colinearidade perfeita, o cálculo de MQO falha porque não é possível separar efeitos de variáveis redundantes. Mais comum, contudo, é a multicolinearidade elevada mas imperfeita, quando variáveis explicativas estão fortemente correlacionadas. Isso não causa viés nos coeficientes, porém infla suas variâncias, tornando as estimativas instáveis e com amplos intervalos de confiança. No nosso exemplo, se incluíssemos simultaneamente "idade" e "anos de experiência" como regressoras, elas estariam altamente correlacionadas. O diagnóstico clássico para multicolinearidade é calcular os fatores de inflação da variância (VIF) para cada coeficiente; valores de VIF acima de 10 são considerados indicativos de multicolinearidade séria. Como solução, costuma-se remover ou combinar variáveis fortemente correlacionadas, ou centrar variáveis (no caso de multicolinearidade envolvendo termos polinomiais, por exemplo). Em nosso modelo de exemplo, poderíamos optar por usar idade e não anos de experiência, ou vice-versa, para evitar redundância excessiva (Stats Made Easy, s.d.). **Exogeneidade das regressoras (ou $E(u|X) = 0$):** Este é talvez o pressuposto conceitualmente mais importante para dar aos coeficientes uma interpretação causal. Exogeneidade significa que o termo de erro não está correlacionado sistematicamente com as variáveis explicativas. Equivalentemente, não há fatores omitidos no erro que sejam correlacionados com X, e não há causalidade reversa de Y para X. Se $E(u|X) = 0$ for atendido, podemos interpretar β_1 como uma estimativa insesgada do efeito ceteris paribus de educação sobre salário. Todavia, se houver endogeneidade – termo usado quando essa suposição falha – os coeficientes de MQO estarão viesados. Por exemplo, a habilidade inerente de um indivíduo pode afetar seus rendimentos e influenciar sua escolaridade; se a habilidade não está medida e incluída em X, ela passa a integrar u_i e, sendo correlacionada com educação, quebra-se a condição $E(u|X) = 0$. Não há teste estatístico perfeito para exogeneidade (pois envolve variáveis não observadas), mas inferimos sua plausibilidade a partir de conhecimento substantivo e teórico. Em nosso caso, é razoável suspeitar que habilidade ou background familiar sejam variáveis omitidas correlacionadas com educação, o que levanta uma bandeira vermelha: ainda que controlemos experiência, gênero e região, a associação educação-salário pode não refletir unicamente um efeito causal da escolaridade. Retornaremos a esse ponto ao discutir endogeneidade e causalidade na próxima seção. Em termos de diagnóstico, para verificar a adequação do modelo e dos pressupostos, usamos uma combinação de gráficos e indicadores numéricos, conforme já insinuado: gráfico de resíduos vs valores ajustados (para checar linearidade e homocedasticidade) (Stats Made Easy, s.d.), gráficos Q-Q dos resíduos (para normalidade), cálculo de estatísticas como teste de Breusch-Pagan (heterocedasticidade) ou Durbin-Watson (autocorrelação, se pertinente), inspeção de outliers via distância de Cook ou alavancagem (que não é um pressuposto em si, mas outliers podem distorcer o ajuste e violar implícitamente suposições) (Stats Made Easy, s.d.). Importante destacar que nenhum modelo real atenderá perfeitamente a todos os pressupostos – o objetivo do diagnóstico é identificar violações severas que possam comprometer a análise e então adotar medidas para mitigá-las ou, em casos extremos, considerar modelos alternativos. Caso os pressupostos não sejam atendidos, as opções incluem: adicionar variáveis explicativas relevantes omitidas, transformar variáveis (como tomar log, adicionar termos quadráticos), utilizar métodos robustos ou não lineares, ou aplicar técnicas específicas (por exemplo, modelos de mínimos quadrados generalizados, estimação robusta, etc.) (Stats Made Easy, s.d.). Em suma, o diagnóstico é parte integrante do método de regressão, garantindo que "lixo não entre e lixo não saia" (rubbish in, rubbish out) (Stats Made Easy, s.d.) – isto é, verificando que o modelo escolhido representa de forma adequada os dados, podemos

confiar mais nas conclusões tiradas.

Interpretação dos Resultados: Coeficientes, Incerteza e Ajuste

Com um modelo ajustado e presumivelmente validado pelos diagnósticos principais, passamos à interpretação substantiva dos resultados. No caso em foco – a relação entre educação e salário (log) – o coeficiente β_1 será o protagonista. Como já mencionado, β_1 representa a diferença percentual aproximada no salário associada a mais um ano de estudo, mantendo-se constantes os demais fatores do modelo. Se $\beta_1 = 0{,}08$, interpretamos que, para dois indivíduos semelhantes em experiência, gênero e região, aquele com um ano a mais de escolaridade ganha em média ˜8% a mais em salário. Essa tradução em termos percentuais facilita a comunicação para um público amplo, pois porcentagens são intuitivas (diferente de dizer "log-pontos de salário"). Os demais coeficientes também têm interpretações ceteris paribus: por exemplo, β_2 associado à experiência talvez seja algo como 0,02, indicando que cada ano adicional de experiência está associado a 2% a mais no salário, tudo o mais constante. O coeficiente de gênero, digamos $\beta_3 = -0{,}15$ tendo feminino como 1 e masculino 0, indicaria que, em média, mulheres ganham cerca de 15% a menos que homens com nível educacional, experiência e região semelhantes – evidência de um possível hiato salarial de gênero. É crucial comunicar esses resultados de forma clara, enfatizando associação condicional em vez de causalidade confirmada, a menos que se esteja muito seguro das hipóteses causais. Junto aos coeficientes, a incerteza estatística deve ser reportada. Tipicamente, fornecemos intervalos de confiança (IC) para cada coeficiente – por exemplo, um IC de 95% para β_1. Se $\beta_1 = 0{,}08$ com IC 95% [0,05; 0,11], isso significa que, considerando a variabilidade amostral, estamos 95% confiantes de que o verdadeiro efeito percentual por ano de estudo está entre 5% e 11%. Intervalos de confiança são geralmente mais informativos que apenas valores-p, pois mostram a gama de efeitos compatíveis com os dados. No exemplo dado, o IC não só sugere que o efeito é positivamente diferente de zero (pois o intervalo não inclui 0), mas também dá ideia da magnitude plausível do retorno da educação, evitando interpretações superprecisas do ponto-estimado. Outra métrica presente em praticamente toda saída de regressão é o coeficiente de determinação R^2, que indica a proporção da variância da variável dependente explicada pelo modelo. Embora R^2 seja útil como medida de ajuste global, deve ser lido com parcimônia, especialmente em ciências sociais. Frequentemente, fenômenos sociais possuem alta variabilidade individual e são influenciados por muitos fatores aleatórios ou não observados, de modo que valores de R^2 da ordem de 0,2 ou 0,3 (20–30% da variância explicada) não são incomuns e ainda assim o modelo pode ter utilidade substantiva. Um equívoco comum de iniciantes é achar que um R^2 baixo invalida o modelo – não necessariamente; significa apenas que há muito que não está sendo explicado, o que é esperado quando se modela comportamento humano. Inversamente, um R^2 muito alto pode ser suspeito: caso encontremos 90% de variação explicada num contexto social, isso pode indicar superajuste (overfitting) ou inclusão de variáveis que são quase a própria variável resposta (por exemplo, incluir salário do ano passado para explicar salário deste ano trivialmente dará R^2 altíssimo). Como ressalta Frost (2023), "comportamento humano possui variabilidade inexplicável muito maior, resultando em valores de R^2 geralmente abaixo de 50%. 90% seria alto demais nesse contexto!" (Frost, 2023). Portanto, ao avaliar R^2, deve-se levar em conta a natureza do problema e comparativos com estudos semelhantes em vez de perseguir cegamente um número alto. Além disso, enfatiza-se a análise de R^2 ajustado (que penaliza número de regressores) para evitar achar que adicionar variáveis irrelevantes "melhorou" o modelo quando apenas inflou o R^2 pela complexidade. Um aspecto frequentemente negligenciado é o escopo de inferência do modelo. Isto refere-se ao a quem e a que contexto nossos resultados se aplicam. Se nossos dados provêm, por exemplo, de uma pesquisa domiciliar nacional de 2020, então as conclusões (como o retorno

de 8% ao ano de escolaridade) aplicam-se àquele país, naquele período, para a população coberta (talvez adultos empregados de 25 a 60 anos, dependendo do desenho amostral). Não podemos supor automaticamente que o mesmo valor valha para outras populações ou épocas – em outro país o retorno pode ser diferente, ou para faixas etárias distintas pode variar. Em termos de inferência, fazemos generalizações do nosso amostra para a população alvo que ela representa. Por isso é importante delinear quem está na amostra: o efeito estimado pode diferir se incluíssemos trabalhadores informais, ou apenas urbanos vs rurais, etc.Além disso, diferenciar associação condicional de causalidade faz parte da validade: nosso modelo controlado pode sugerir que educação tem associação positiva com salário independentemente de outros fatores observados, mas ainda assim não prova causalidade, a menos que certas condições sejam atendidas (exogeneidade, ou experimentação aleatória, etc.). Portanto, na interpretação final comunicada, deve-se evitar linguagem determinística do tipo "um ano de educação garante 8% a mais de salário" e preferir "está associado, em média, a 8% a mais, controlados fatores X, Y, Z". Assim, deixamos claro o nível de suposição envolvido. Resumindo esta parte, interpretar um modelo linear requer traduzir números em linguagem substantiva acessível, quantificar incertezas e reconhecer os limites do que aqueles números podem afirmar sobre a realidade. Especialmente em ciências sociais, transparência na comunicação das suposições e do alcance dos resultados é fundamental para evitar leituras enganosas ou supergeneralizações.

Limites do Controle Multivariado: Endogeneidade e Causalidade

Por mais útil que seja a regressão múltipla para revelar associações condicionais, ela possui limites notáveis quando o objetivo é inferir causalidade. O caso educação $\rightarrow$ salário é emblemático. Se quisermos interpretar β_1 como o efeito causal de mais um ano de estudo no salário futuro de um indivíduo, precisamos que não haja viés de variável omitida ou outros problemas de endogeneidade. Entretanto, há fortes razões para crer que indivíduos com mais escolaridade diferem dos menos escolarizados em aspectos não totalmente captados pelos controles usuais. Dois candidatos frequentemente mencionados são a habilidade (talento) inato e o background familiar socioeconômico. Indivíduos mais habilidosos podem tanto obter mais anos de estudo (por conseguirem melhor desempenho acadêmico, oportunidades educacionais, etc.) quanto ter maiores salários de qualquer forma (por serem mais produtivos ou ambiciosos) (IZA World of Labor, s.d.). Assim, quando comparamos pessoas com 12 vs 11 anos de estudo na nossa regressão, talvez a de 12 anos em média seja também aquela com um pouco mais de habilidade, inflando seu salário independentemente da educação. Neste caso, atribuir toda a diferença salarial à educação superestimaria o verdadeiro efeito causal – estaríamos na verdade capturando o "prêmio" da habilidade também. Esse fenômeno é exatamente o viés de variável omitida (VVO): ocorre quando um fator não observado influencia tanto a variável explicativa quanto a dependente, levando a uma correlação espúria entre elas (IZA World of Labor, s.d.). Formalmente, se habilidade está ausente do modelo e $\text{Corr}(\text{educação}, \text{habilidade}) \neq 0$ e habilidade também afeta salário, então a exogeneidade falha e $\hat{\beta}_1$ do MQO será viesado. No exemplo, supõe-se que a correlação de educação com habilidade seja positiva e habilidade eleva salários; isso implicaria viés positivo – nosso coeficiente de educação estaria medindo "efeito da educação + algo do efeito da habilidade". Por isso, algumas pesquisas que não consideram bem essa questão podem encontrar "retornos da educação" inflados. Com efeito, estimativas ingênuas via MQO sugerem algo em torno de 6–10% de aumento salarial por ano de estudo (IZA World of Labor, s.d.), mas parte desse valor pode refletir não retorno causal puro, e sim diferenças preexistentes entre quem estuda mais ou menos. Consequentemente, "OLS não seria informativo sobre o efeito de uma política que aumente os anos de educação" se a endogeneidade

estiver presente (IZA World of Labor, s.d.), já que a política afetaria também pessoas de habilidade distinta. Reconhecer esses limites não invalida a regressão múltipla, mas nos força a contextualizar conclusões: nosso coeficiente de 8% pode ser interpretado como uma associação condicional ou, no jargão de Lundberg et al. (2021), como um estimador empírico que se conecta a um estimando teórico sob certas hipóteses (Lundberg; Johnson; Stewart, 2021). Quais hipóteses? Essencialmente, de que controlamos tudo que importa (além de não errar na forma funcional etc.). Se admitimos que não controlamos habilidade ou outros não observados, então falta cumprir uma condição de identificação para interpretar causalmente (Lundberg; Johnson; Stewart, 2021). Em contrapartida, se tivéssemos uma forma de "quebrar" essa correlação espúria, poderíamos estimar o efeito causal de maneira mais confiável. Uma das soluções conceituais para o problema de endogeneidade é recorrer a variáveis instrumentais (VI). Um instrumento é uma variável Z correlacionada com a variável explicativa de interesse (X, educação) mas não correlacionada com os fatores não observados do erro, afetando Y (salário) somente através de X. Encontrar um instrumento válido é notoriamente difícil, mas no caso de educação e rendimentos, pesquisadores ao longo do tempo identificaram alguns candidatos engenhosos. Por exemplo, mudanças na legislação educacional que obrigam coortes de estudantes a permanecer um ano a mais na escola servem como "experimentos naturais". Se, digamos, em certo ano a lei passou a exigir 9 anos mínimos de estudo em vez de 8, isso força alguns indivíduos que teriam saído com 8 anos a estudarem 9 – efetivamente introduzindo variação em educação independente da vontade ou habilidade individual. De fato, um estudo hipotético poderia comparar salários futuros de pessoas que, por um timing de nascimento, foram obrigadas a uma escolaridade extra, com as imediatamente mais velhas que não foram, presumindo que ambos os grupos são similares em tudo mais (pois a mudança foi exógena) (IZA World of Labor, s.d.). Essa diferença de salários (adequadamente analisada) forneceria uma estimativa causal do retorno daquele ano extra de educação. Instrumentos clássicos usados em pesquisas reais incluem: trimestre de nascimento, distância até a escola ou faculdade mais próxima, loterias de serviço militar que afetam incentivo a estudar, entre outros (IZA World of Labor, s.d.).

BOX Um bom instrumento é relevante (correlaciona-se com X) e exógeno (não correlaciona-se com u). Ele não elimina a endogeneidade estrutural, mas isola variação exógena em X que pode ser usada para inferir causalidade.

Não entraremos em detalhes matemáticos de como a estimativa por VI (como a técnica de dois estágios, 2SLS) é feita, pois isso foge do escopo introdutório. O ponto central é motivar que, quando o controle multivariado usual falha em eliminar viés (por conta de variáveis inobserváveis), métodos alternativos focados em desenho de pesquisa (experimentos, quase-experimentos) ou técnicas como variáveis instrumentais são necessários para alcançar inferências causais confiáveis. Em outras palavras, a regressão linear tradicional, sozinha, tem limites para afirmar causalidade: ela necessita de suporte de suposições não testáveis (como "não há omitidos correlacionados") ou de estratégias de identificação complementares. Por isso, em cursos futuros, temas como regressão instrumental, diferenças-em-diferenças, modelos de efeitos fixos e regressão descontínua ganham destaque – todos buscando contornar a endogeneidade de formas diferentes (Angrist; Pischke, 2009; Morgan; Winship, 2015). No caso pedagógico que acompanhamos, fica a lição: nosso estimador MQO de 8% ao ano provavelmente superestima o efeito causal verdadeiro da educação sobre salários devido ao viés de habilidade omitida (IZA World of Labor, s.d.). Estudos usando variáveis instrumentais muitas vezes encontram retornos menores, digamos na faixa de 5–8%, sugerindo que, descontado o viés, esse seria o efeito mais "puro" da educação adicional. Contudo, instrumentos podem introduzir outras sutilezas (por exemplo, estimar um efeito local para quem é afetado pelo instrumento). O importante é que

reconheçamos os limites do controle observacional: nem sempre adicionar controles resolve – quando fatores cruciais não podem ser medidos, é preciso recorrer a outros delineamentos ou aceitar a incerteza quanto à causalidade.

Fio Condutor e Síntese Integrada: Educação e Rendimentos

Recapitulando de forma integrada, acompanhamos o percurso de uma análise hipotética da relação entre educação e salário, desde a descrição inicial até as considerações causais finais. Começamos posicionando a questão no contexto maior da análise multivariada e definindo claramente nosso estimando – o retorno médio da educação. Após um exame exploratório indicando correlação positiva entre anos de estudo e salários, formulamos um modelo de regressão linear simples que quantificou essa tendência (por exemplo, retornos ˜10% anuais). Em seguida, evoluímos para um modelo múltiplo incluindo variáveis de controle relevantes (experiência, gênero, região), o que refinou a estimativa do coeficiente de educação (por exemplo, reduzindo-o para ˜8%) ao isolar melhor o efeito educacional, evidenciando o conceito de efeitos condicionais. No processo, enfatizamos os pressupostos necessários para que esse modelo fosse válido: verificamos graficamente a linearidade (a relação log-salário vs educação parecia aproximadamente reta), não detectamos padrões claros de heterocedasticidade gritante (mas por segurança cogitamos usar erros robustos), observamos que os resíduos seguiam distribuição aproximadamente normal (dando confiabilidade aos intervalos de confiança calculados) e notamos que nossas variáveis de controle não eram multicolineares a ponto de impedir a estimação (embora experiência e idade fossem relacionadas, optamos por incluir apenas uma delas). Essas checagens diagnósticas fizeram parte do fluxo da análise, garantindo que não ignoramos os sinais de alerta que os dados poderiam fornecer. Ao interpretar os resultados do modelo final, traduzimos coeficientes em implicações substantivas: um ano extra de estudo associou-se a ˜8% de salário maior, controlados os fatores mencionados; mulheres ganhavam X% a menos que homens equivalentes; cada ano de experiência rendia Y% de aumento salarial, etc. Também ressaltamos a incerteza inerente – por exemplo, o retorno de 8% vinha com um intervalo plausível de, digamos, 5 a 11%. O ajuste geral do modelo, medido por R^2, foi moderado (supostamente 0,3, i.e. 30% da variação do log-salário explicada), o que julgamos razoável para dados individuais em ciências sociais, alertando para não supervalorizar uma busca por R^2 alto, e sim focar na validade e significância dos coeficientes principais. Por fim, discutimos as limitações e implicações causais: reconhecemos que, apesar de todos os controles, a análise multivariada convencional pode falhar em estabelecer causalidade se fatores como habilidade não estão presentes nos dados. Explicitamos que o viés de variável omitida (aqui, habilidade ou background familiar) provavelmente contamina a estimativa de retorno da educação. Isso nos levou a introduzir conceitualmente a ideia de variáveis instrumentais, mostrando como um desenho de pesquisa diferente (por exemplo, uma mudança de legislação educacional) poderia fornecer evidências mais próximas de causais. Essa discussão cumpre um papel didático de situar o que aprendemos (regressão linear) dentro de um espectro mais amplo de métodos: sabemos agora o que ela faz bem (estimar associações condicionais, detectar relações lineares, fazer predições em âmbito exploratório) e onde devemos ter cautela (afirmações de causalidade). Notemos que, em cada etapa, tomamos o cuidado de explicitar as escolhas de modelagem, as suposições envolvidas e o público-alvo das inferências. Desde o início, imaginamos que nossa população de interesse fossem trabalhadores adultos no mercado formal; deixamos subentendido (mas convém sempre registrar) que nossa inferência vale para esse grupo. O estimando implícito – o efeito médio condicional da educação sobre salário – foi tratado com rigor conceitual: é uma diferença média, não uma certeza determinística para cada indivíduo, e válida sob

condições. Esse hábito de clareza torna a pesquisa mais transparente e reprodutível.

Comunicação dos Resultados e Boas Práticas

Um aspecto transversal a todo trabalho científico – e que merece ênfase especial ao fim de um estudo – é como comunicar os resultados de maneira honesta, transparente e útil. Uma análise de regressão linear, bem conduzida, deve vir acompanhada de informações que permitam a outros entenderem exatamente o que foi feito, quais as premissas adotadas e quão confiáveis são as conclusões. Vamos pontuar algumas boas práticas de comunicação e "governança mínima" de resultados: **Transparência de suposições:** Deixe claro no relato quais pressupostos do modelo foram verificados e quais permanecem como suposições não testadas. Por exemplo, pode-se afirmar: "Verificamos que a relação entre as variáveis parecia linear e que os resíduos não apresentaram heterocedasticidade evidente; assumimos que não há fatores omitidos correlacionados com escolaridade, embora reconheçamos essa como uma limitação." Essa frase comunica tanto o que foi checado (linearidade, homocedasticidade) quanto o que é assumido (exogeneidade), dando ao leitor insumos para julgar a credibilidade da análise. **Relato de diagnósticos:** Em complemento ao ponto anterior, inclua na comunicação menção a quaisquer problemas detectados e como foram tratados. Exemplo: "Como houve leve indício de heterocedasticidade, reportamos intervalos de confiança com erros-padrão robustos." Ou: "Dois pontos de alavancagem alta foram identificados, mas a remoção deles não alterou substantivamente os coeficientes, então optamos por mantê-los." Esse tipo de transparência aumenta a confiança de que o pesquisador inspecionou possíveis distorções. **Escopo de inferência:** Reitere a que população e condição se referem os resultados. "Esses resultados referem-se a trabalhadores formais de 25-60 anos no Brasil em 2020, baseados em dados da PNAD contínua; extrapolações para outros grupos devem ser feitas com cautela." Informação assim contextualiza o leitor e impede interpretações abusivas (por exemplo, achar que 8% valha para qualquer país ou período). **Linguagem precisa sobre causalidade:** Se não foi possível eliminar totalmente dúvidas de endogeneidade, evite termos como "impacto" ou "efeito causal" no resumo dos achados. Prefira "associação" ou no máximo "efeito estimado, assumindo que controlamos adequadamente outros fatores". Por outro lado, se há forte convicção causal (por razões externas ou desenho quase-experimental), explique por quê. No nosso caso, provavelmente enfatizaríamos que não podemos garantir causalidade devido à possibilidade de viés de omitidos, mas que o achado é consistente com a ideia de retorno econômico da educação. **Reprodutibilidade básica:** Sempre que possível, mencione a fonte dos dados e a disponibilidade de códigos ou procedimentos analíticos. Em um relatório acadêmico, isso poderia significar disponibilizar um apêndice metodológico ou link para um repositório contendo o script (R, Python, Stata etc.) usado na regressão. No contexto de aula, pode-se ao menos descrever quais passos foram seguidos para que outros possam replicar (por exemplo: "Foi utilizada regressão linear pelo método MQO; variáveis X, Y, Z foram transformadas conforme explicado; todos os cálculos foram feitos no software R"). Tais práticas são cada vez mais valorizadas na ciência social quantitativa atual para garantir confiabilidade e permitir a outros aprenderem diretamente recriando a análise. **Governança e ética na análise:** Isso envolve reconhecer limites e conflitos de interesse. No nosso caso, poderia não se aplicar muito, mas imagine comunicar a um formulador de políticas: seria importante frisar que simplesmente aumentar anos de escolaridade por decreto não garante automaticamente aumentos salariais proporcionais se a qualidade da educação não for boa (um ponto frequentemente discutido na literatura de economia da educação). Ou seja, incluir contexto qualitativo e teórico na comunicação: números não falam sozinhos; o pesquisador deve guiá-los com interpretações informadas. Em suma, a fase final de qualquer projeto de regressão linear em ciências sociais conecta os resultados numéricos de volta às

perguntas teóricas ou de política pública originais, respeitando os achados mas também delineando limitações. No exemplo de educação e rendimentos, a comunicação responsável dos resultados resumiria algo como: "Encontramos que, em média, um ano adicional de estudo associa-se a ˜8% mais de salário entre trabalhadores brasileiros, controlando experiência, gênero e região (IC 95% aproximadamente 5%–11%). Esse valor sugere retornos econômicos substanciais da educação, embora não possamos afirmar que se trata de efeito puramente causal – pessoas com mais educação podem diferir em outros atributos não observados, como habilidade. Ainda assim, o padrão é consistente com a literatura de capital humano. Os resultados realçam a importância da educação na determinação de renda, mas também apontam para a necessidade de cautela: políticas educacionais que elevem a escolaridade provavelmente terão impacto positivo nos rendimentos, mas a magnitude exata desse impacto pode variar. Todos os dados utilizados são da PNAD, e análises detalhadas (incluindo checagem de pressupostos e códigos) estão documentadas no apêndice para escrutínio público." Uma mensagem como essa amarra o estimando teórico à evidência empírica, com rigor e humildade, que é o objetivo final de uma análise quantitativa bem conduzida. E assim encerramos este percurso pelos métodos de regressão linear aplicada às ciências sociais, tendo o caso "educação e salário" como guia prático para compreender conceitos, calcular modelos, diagnosticar problemas, interpretar números e reconhecer até onde podemos ir com as ferramentas aprendidas e onde precisamos recorrer a próximas etapas de aprofundamento metodológico.

Referências Bibliográficas

ANGRIST, Joshua D.; KRUEGER, Alan B. Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, v. 106, n. 4, p. 979–1014, 1991.

ANGRIST, Joshua D.; PISCHKE, Jörn-Steffen. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press, 2009.

ATHEY, Susan; IMBENS, Guido W. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, v. 113, n. 27, p. 7353–7360, 2016.

BRAND, Jennie E.; ZHOU, Xiang; XIE, Yu. Recent developments in causal inference and machine learning. *Annual Review of Sociology*, v. 49, p. 81–110, 2023.

BRADY, Henry E. Causation and explanation in social science. In: BOX-STEFFENSMEIER, Janet M.; BRADY, Henry E.; COLLIER, David (org.). *The Oxford Handbook of Political Methodology*. New York: Oxford University Press, 2008.

BREIMAN, Leo. Statistical modeling: the two cultures. *Statistical Science*, v. 16, n. 3, p. 199–231, 2001.

CHATTOPADHYAY, Ambarish; ZUBIZARRETA, José R. Causation, comparison, and regression. *Harvard Data Science Review*, v. 6, n. 1, 2024.

CHERNOZHUKOV, Victor; CHETVERIKOV, Denis; DEMIRER, Mert; DUFLO, Esther; HANSEN, Christian; NEWEY, Whitney K.; ROBINS, James M. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, v. 21, n. 1, p. C1–C68, 2018.

CUNNINGHAM, Scott. *Causal Inference: The Mixtape*. New Haven: Yale University Press, 2021.

GRIMMER, Justin. We are all social scientists now: how big data, machine learning, and causal inference work together. *PS: Political Science & Politics*, v. 48, n. 1, p. 80–83, 2015.

IMBENS, Guido W.; ANGRIST, Joshua D. Identification and estimation of local average treatment effects. *Econometrica*, v. 62, n. 2, p. 467–475, 1994.

LUNDBERG, Ian; JOHNSON, Rebecca; STEWART, Brandon M. What is your estimand? *American Sociological Review*, v. 86, n. 3, p. 532–565, 2021.

MORGAN, Stephen L.; WINSHIP, Christopher. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. 2. ed. New York: Cambridge University Press, 2015.

PEARL, Judea. The foundations of causal inference. *Sociological Methodology*, v. 40, n. 1, p. 75–149, 2010.

SALGANIK, Matthew J. *Bit by Bit: Social Research in the Digital Age*. Princeton: Princeton University Press, 2019.

SHMUELI, Galit. To explain or to predict? *Statistical Science*, v. 25, n. 3, p. 289–310, 2010.

ANGRIST, Joshua D.; KRUEGER, Alan B. Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, v. 106, n. 4, p. 979–1014, 1991.

ANGRIST, Joshua D.; PISCHKE, Jörn-Steffen. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press, 2009.

FOX, John. *Applied Regression Analysis and Generalized Linear Models*. 3. ed. Thousand Oaks: SAGE, 2016.

FROST, Jim. Five reasons why your R-squared can be too high. *Statistics by Jim* (blog), 2023. Disponível em: https://statisticsbyjim.com/regression/r-squared-too-high/. Acesso em: dia mês ano.

GELMAN, Andrew; HILL, Jennifer. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press, 2007.

HAIR, Joseph F.; BLACK, William C.; BABIB, Barry J.; ANDERSON, Rolph E.; TATHAM, Ronald L. *Multivariate Data Analysis*. 8. ed. Andover: Cengage Learning, 2019.

IMBENS, Guido W.; ANGRIST, Joshua D. Identification and estimation of local average treatment effects. *Econometrica*, v. 62, n. 2, p. 467–475, 1994.

LEWIS-BECK, Colin; LEWIS-BECK, Michael. *Applied Regression: An Introduction*. 2. ed. Thousand Oaks: SAGE, 2015.

LUNDBERG, Ian; JOHNSON, Rebecca; STEWART, Brandon M. What is your estimand? *American Sociological Review*, v. 86, n. 3, p. 532–565, 2021.

WOOLDRIDGE, Jeffrey M. *Introductory Econometrics: A Modern Approach*. 7. ed. Boston: Cengage Learning, 2020.