

Viés de Variável Omitida

IBM0288 - 2026.2

Prof. Raphael Gouvea

1 Para reflexão



Quentin Gallea, PhD • Seguindo

Causal AI Training, Advisory and Keynotes | Helping Measure your A...
2 sem •

I would not say that human expertise "still" matters in the age of AI, but rather that it matters more than ever.
Take my field of expertise, causal inference. Before people could not run a analysis on their own, they didn't know the model to apply, the code, or the technical terms.

However, if you ask GenAI, you'll get a fully functioning code, explanation, interpretation of the results in minutes. No bugs and a lot of confidence, and with almost certainty, it will be completely wrong...

Companies, advisors, data scientists, approached me to show what they do and confirm that it works. I can tell you that I've seen tons of absolute non-sense (even with product already commercialized...).

In a field like causal inference, you do need very deep expertise to do it well and manage the uncertainty. And with more powerful model, you need even more expertise because these models can go deeper, and further alone. So in order to really guide (filter-in) and check (filter-out) you need the expertise.

Bottom line, I do think that deep expertise will be (should be) more valued than ever.

The problem is that many people like "confident lies"...

2 Aula passada: regressões e causalidade

- Quando β_1 pode ser interpretado como efeito causal médio de X sobre Y ?
- X precisa ser independente de outros fatores que afetam Y
- Isso acontece para dados experimentais!
- Não será sempre verdadeiro para dados observacionais!

3 Aula passada: 3 hipóteses para inferência causal

+ A variável explicativa X é independente do termo de erro u_i

$$E[u_i | X_i] = 0; \text{corr}(X_i, u_i) = 0$$

2. $(X_i, Y_i), i = 1, \dots, n$, são independentes e identicamente distribuídos (i.i.d.).

- Algumas violações de independência podem ser resolvidas com séries temporais ou dados em painel
3. Sem grandes outliers em X e/ou Y .
- A presença de outliers pode distorcer os estimadores de MQO.

4 Por que incluir mais variáveis no modelo?

- Se o objetivo for fazer previsão, incluir variáveis pode aumentar a precisão das previsões.
- Se o objetivo for inferência causal, incluir variáveis permite controlar por outros determinantes da variável dependente que levariam a viés.

...

Aviso

Incluir variáveis aleatoriamente no modelo não é uma boa ideia! A inclusão de controles ruins pode introduzir viés ao invés de eliminá-lo e inviabilizar a identificação de efeitos causais.

5 Viés de Variável Omitida

O viés de variável omitida (OVV) ocorre quando:

1. A variável omitida afeta a variável dependente Y , ou seja, ela é parte de u .
2. A variável omitida é correlacionada com o regressor X incluído.

6 OVB no modelo de notas e tamanho das turmas

$$\text{TestScore}_i = \beta_0 + \beta_1 \text{STR}_i + u_i$$

A omissão de alguma das seguintes variáveis gera viés no modelo?

1. Capacidade financeira do distrito escolar
2. Horário do teste
3. Tamanho médio do estacionamento dos professores nas escolas do distrito.
4. Percentual de alunos para os quais inglês não é o primeiro idioma ($PctEL$)

7 Experimento ideal

- Imagine um experimento controlado randomizado para medir o efeito dos tamanhos das turmas sobre a nota do teste padronizado.

- Nesse experimento, os alunos seriam designados aleatoriamente para turmas que teriam tamanhos diferentes.
- Como a atribuição nas turmas seria aleatorizada, todas as características dos alunos seriam distribuídas independentemente do tamanho das turmas
- Assim, $E(u_i | \text{STR}_i) = 0$: ou seja, a hipótese 1 acima seria válida.

8 Como dados observacionais diferem desse ideal?

- O tratamento não é atribuído aleatoriamente.

- Considere o percentual de alunos para os quais inglês não é o primeiro idioma:
 - plausivelmente satisfaz os dois critérios para viés de variável omitida
- Assim, os grupos de “controle” e “tratamento” diferem de forma sistemática, de modo que $\text{corr}(STR, PctEL) \neq 0$

...

- Examinando o efeito do tamanho das turmas entre distritos com mesmo $PctEL$ eliminamos a diferença entre controle (turma grande) e tratamento (turmas pequenas).

- Se a única diferença sistemática entre os grupos fosse essa, estaríamos de volta ao experimento controlado randomizado ao fazer a análise dentro de cada grupo de $PctEL$.
- Essa é uma forma de **controlar** o efeito do $PctEL$ ao estimar o efeito do STR .
 - a outra é fazer uma regressão incluindo $PctEL$ como controle.

9 OVB: expressão matemática

Seja β_1 o verdadeiro efeito causal de X sobre Y na população e $\rho_{Xu} = \text{corr}(X_i, u_i)$.

Sob essas hipóteses, o coeficiente de MQO será:¹

$$E[\hat{\beta}_1] = \beta_1 + \rho_{Xu} \frac{\sigma_u}{\sigma_X}$$

...

! Importante

O tamanho da amostra não resolve o problema do viés de variável omitida. O estimador $\hat{\beta}_1$ é viesado e inconsistente quando há OVB.

10 OVB: modelo longo vs curto

Regressão Longa: $Y_i = \alpha^l + \beta^l X_i + \gamma Z_i + e_i^l$

Regressão Curta: $Y_i = \alpha^c + \beta^c X_i + e_i^c$

! Efeito tratamento e OVB

A estimativa da regressão curta é igual à da regressão longa mais o efeito da omitida vezes o coeficiente da regressão da omitida na incluída. Ou seja, a regressão curta estima o **efeito causal** mais o **viés**.

¹Demonstração nos Apêndices 4.3 e 6.1 (4ª edição) e 4.3 e 5.1 (1ª edição) de Stock e Watson.

11 Derivação

$$\begin{aligned}
 \beta^c &= \frac{Cov(Y_i, X_i)}{Var(X_i)} \\
 &= \frac{Cov(\alpha^l + \beta^l X_i + \gamma Z_i + e_i^l, X_i)}{Var(X_i)} \\
 &= \frac{\beta^l Var(X_i) + \gamma Cov(Z_i, X_i) + Cov(e_i^l, X_i)}{Var(X_i)} \\
 &= \beta^l + \gamma \frac{Cov(Z_i, X_i)}{Var(X_i)} \\
 &= \beta^l + \gamma \pi_{21}
 \end{aligned}$$

onde π_{21} é o coeficiente de X_i em $Z_i = \pi_{20} + \pi_{21}X_i + u_i$.

12 Qual o sinal esperado do OVB?

Seja: Y a variável dependente, X a variável independente e Z a variável omitida.

A equação para o viés é dada por: $E[\hat{\beta}^l] = \beta^l + \gamma \pi_{21}$

$$\text{Corr}(Z, X) > 0 \quad \text{Corr}(Z, X) < 0$$

$$\text{Corr}(Z, Y) > 0$$

$$\text{Corr}(Z, Y) < 0$$

$$\text{Corr}(Z, X) > 0 \quad \text{Corr}(Z, X) < 0$$

$$\text{Corr}(Z, Y) > 0 \text{ viés p/ cima } \uparrow$$

$$\text{Corr}(Z, Y) < 0$$

$$\text{Corr}(Z, X) > 0 \quad \text{Corr}(Z, X) < 0$$

$$\text{Corr}(Z, Y) > 0 \text{ viés p/ cima } \uparrow \quad \text{viés p/ baixo } \downarrow$$

$$\text{Corr}(Z, Y) < 0$$

$$\text{Corr}(Z, X) > 0 \quad \text{Corr}(Z, X) < 0$$

$$\text{Corr}(Z, Y) > 0 \text{ viés p/ cima } \uparrow \quad \text{viés p/ baixo } \downarrow$$

$$\text{Corr}(Z, Y) < 0 \text{ viés p/ baixo } \downarrow$$

$$\text{Corr}(Z, X) > 0 \quad \text{Corr}(Z, X) < 0$$

$$\text{Corr}(Z, Y) > 0 \text{ viés p/ cima } \uparrow \quad \text{viés p/ baixo } \downarrow$$

$$\text{Corr}(Z, Y) < 0 \text{ viés p/ baixo } \downarrow \quad \text{viés p/ cima } \uparrow$$

13 Solução 1: aleatorização

No caso de experimentos aleatórios:

- $E(X)$ igual para todo valor de X , independente de outros fatores que afetam Y .

- $E(u)$ não varia com $X \rightarrow \text{corr}(X_i, u_i) = 0$.

...

Portanto: a aleatorização de X elimina viés de variável omitida (e causalidade reversa).

14 Solução 2: controlando pelas variáveis omitidas

No caso de dados observacionais:

- Não há garantia de $\text{corr}(X, u) = 0$.
- Mas se pudermos **observar as variáveis omitidas** que afetam tanto Y quanto X , podemos controlar por elas.
- Comparar Y entre unidades com níveis similares de Z mas diferentes níveis de X .
- O que isso significa intuitivamente?

15 Intuição: inglês como 2º idioma

TABLE 6.1 Differences in Test Scores for California School Districts with Low and High Student-Teacher Ratios, by the Percentage of English Learners in the District						
	Student-Teacher Ratio < 20		Student-Teacher Ratio ≥ 20		Difference in Test Scores, Low vs. High Student-Teacher Ratio	
	Average Test Score	n	Average Test Score	n	Difference	t -statistic
All districts	6574	238	650.0	182	74	4.04
Percentage of English learners						
< 1.9%	664.5	76	665.4	27	-0.9	-0.30
1.9-8.8%	665.2	64	661.8	44	3.3	1.13
8.8-23.0%	654.9	54	649.7	50	5.2	1.72
> 23.0%	636.7	44	634.8	61	1.9	0.68

O percentual de alunos para os quais o inglês é o segundo idioma é positivamente correlacionado com o tamanho médio das turmas!

Qual a direção do viés? Você espera um β_1 maior ou menor?

16 MQO com 2 regressores

Função de expectativa condicional:

$$E[Y_i | X_{1i}, X_{2i}] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

Modelo de regressão:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

17 Como interpretar β_1 ?

- $\beta_1 = \frac{\Delta Y}{\Delta X_1}$, mantendo X_2 constante.

- Efeito parcial de X_1
- Como interpretar β_2 ? E β_0 ?

18 MQO com k regressores:

Função de expectativa condicional:

$$E[Y_i | X_{1i}, X_{2i}] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki}$$

Modelo de regressão:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + u_i$$

19 Estimação: estratégia MQO

Selecionar $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ que melhor ajustem os dados.

Melhor ajuste = minimizar os erros quadrados de previsão:

$$\arg \min_{b_0, b_1, \dots, b_k} \sum_{i=1}^n (Y_i - [b_0 + b_1 X_{1i} + b_2 X_{2i} + \dots + b_k X_{ki}])^2$$

20 Os estimadores de MQO

- Modelo de regressão linear: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \dots + \hat{\beta}_k X_{ki} + u_i$
- Os coeficientes de MQO $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots$ e $\hat{\beta}_k$ **estimam** os parâmetros populacionais $\beta_0, \beta_1, \beta_2, \dots$ e β_k a partir dos dados da amostra
- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \dots + \hat{\beta}_k X_{ki}$: valor previsto de Y_i com base em X_i
- $\hat{u}_i = Y_i - \hat{Y}_i$: resíduo (estimador do erro u_i)

21 O teorema Frisch-Waugh-Lovell

Com um regressor ($Y_i = \beta_0 + \beta_1 X_{1i} + u_i$):

$$\hat{\beta}_1 = \frac{cov(X, Y)}{var(X)}$$

Com múltiplos regressores ($Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + u_i$):

$$\hat{\beta}_1 = \frac{cov(\tilde{X}_1, \tilde{Y})}{var(\tilde{X}_1)}$$

$\tilde{X}_1$: resíduos da regressão de X_1 em todos os outros regressores ($X_2, X_3, \dots, X_k$).

$\tilde{Y}$: resíduos da regressão de Y em todos os outros regressores ($X_2, X_3, \dots, X_k$).

22 Estimando β 's em três passos

O teorema Frisch-Waugh-Lovell significa que β_1 pode ser estimado em 3 passos:

1. Regredir X_1 em $X_2, X_3, \dots, X_k$ e obter os resíduos $\tilde{X}_1$
 2. Regredir Y em $X_2, X_3, \dots, X_k$ e obter os resíduos $\tilde{Y}$
 3. Regredir $\tilde{Y}$ em $\tilde{X}_1$
- $\tilde{Y}_i = \beta_0 + \beta_1 \tilde{X}_1 + u_i$

23 OVB no exemplo

Modelo com um regressor:

$$\widehat{TestScore} = 698.9 - 2.28 \text{ STR}$$

Modelo com controle:

$$\widehat{TestScore} = 686.0 - 1.10 \text{ STR} - 0.65 \text{ PctEL}$$

O que aconteceu com o valor do coeficiente de STR? Por quê?

24 Relembrando o sinal do viés

$$\begin{array}{l} \text{Corr}(Z, X) > 0 \quad \text{Corr}(Z, X) < 0 \\ \text{Corr}(Z, Y) > 0 \quad \text{viés p/ cima} \uparrow \quad \text{viés p/ baixo} \downarrow \\ \text{Corr}(Z, Y) < 0 \quad \text{viés p/ baixo} \downarrow \quad \text{viés p/ cima} \uparrow \end{array}$$

25 R^2 e R^2 ajustado

- $R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 - \frac{SSR}{TSS}$
- R^2 sempre aumenta com a inclusão de um regressor
- R^2 ajustado ou $\bar{R}^2 = 1 - \frac{n-1}{n-k-1} \frac{SSR}{TSS}$

...

SSR: Soma dos quadrados dos resíduos

26 SER (erro-padrão da regressão)

$$SER = \hat{\sigma}_u = \sqrt{\frac{1}{n-k-1} \sum_{i=1}^n \hat{u}_i^2}$$

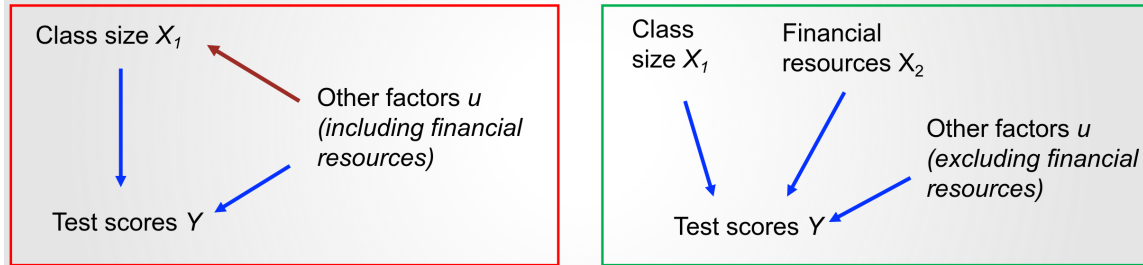
- Mede a dispersão dos resíduos em torno da reta.
- $n - k - 1$ é chamado de graus de liberdade: um ajuste pelo viés introduzido pelo número de parâmetros estimados

27 Hipóteses para inferência causal

1. Os regressores X_s são independentes do erro u_i : $E(u_i | X_{1i}, X_{2i}, \dots, X_{ki}) = 0$

2. $(Y_i, X_{1i}, X_{2i} \dots X_{ki}), i = 1, \dots, n$, são independentes e identicamente distribuídos (i.i.d.).
3. Sem outliers relevantes.
4. Não há multicolinearidade perfeita.

28 Exemplo no modelo das notas e tamanho das turmas



Exemplo hipotético: o tamanho das turmas X_1 não é correlacionado com o termo de erro somente após controlar pelos recursos financeiros do distrito X_2

29 Hipótese de independência condicional

- Seja X o regressor ou tratamento de interesse
- $W_1, W_2, \dots, W_k$ são as variáveis de controle
- Hipótese de independência condicional (CIA):

$$E(u_i | X, W_1, W_2, \dots, W_3) = E(u_i | W_1, W_2, \dots, W_3)$$

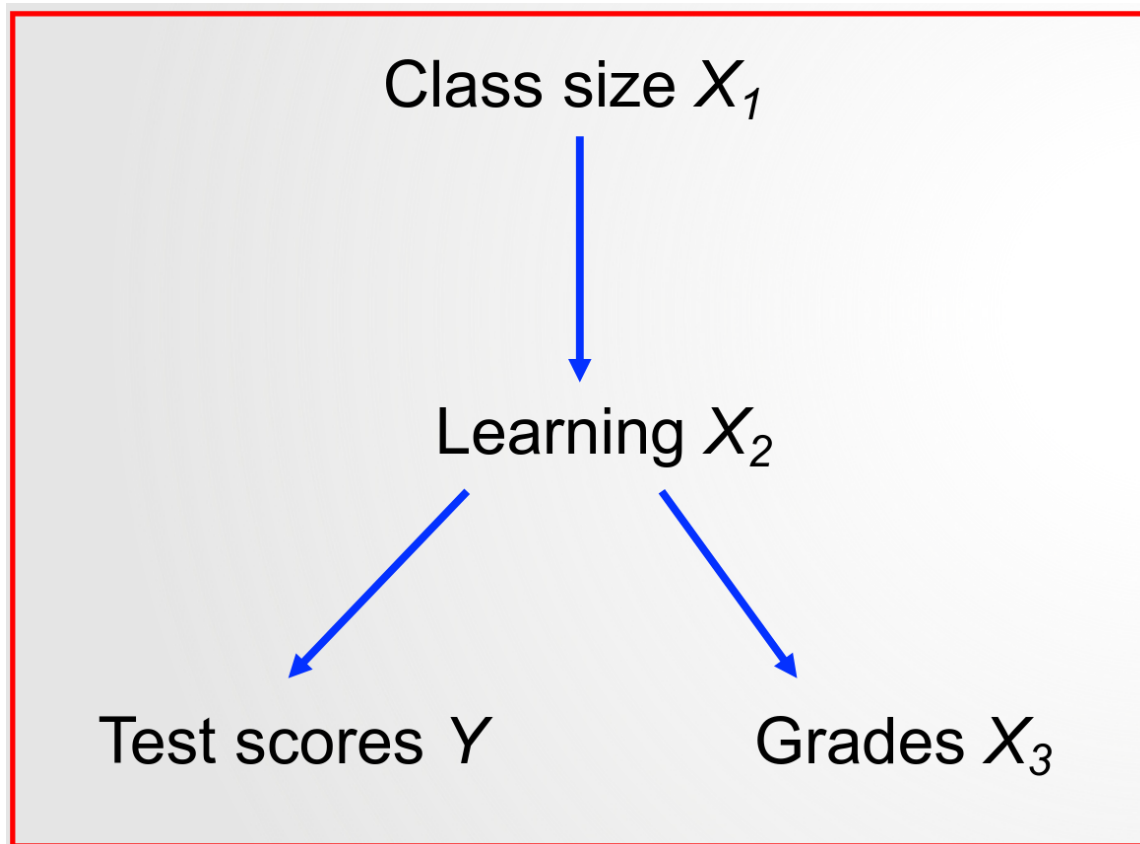
...

Resumo: u e X não são correlacionados após controlar pelos W 's.

30 Controles bons e ruins

- Nem toda variável é uma boa opção para ser incluída como controle
- **Controles Ruins:** variáveis que são afetadas pela variável de interesse X
 - ao mantê-las fixas, introduz-se viés no modelo
- **Controles Bons:** variáveis pré-determinadas em relação à variável de interesse

31 Exemplo de controle ruim



- Objetivo é estimar o efeito tratamento do tamanho das turmas nas notas em exames padronizados (test scores)

- Não controlar por aprendizado! Nem por notas regulares (grades), já que são afetadas por aprendizado!
 - Não faz sentido mantê-las fixas quando o tamanho da sala as afeta.
- Aprendizado e notas regulares são controles ruins!

...

! Importante

Não controle por nada que seja afetado pelo regressor de interesse!

32 Evite multicolinearidade perfeita

$$\widehat{TestScore}_i = \beta_0 + \beta_1 STR_i + \beta_2 PctEL_i + \beta_3 FracEL_i + u_i$$

- $PctEL_i$: percentual de alunos que estão aprendendo inglês (entre 0 e 100)
- $FracEL_i$: fração de alunos que estão aprendendo inglês (entre 0 e 1)
- Multicolinearidade perfeita: exemplo $PctEL = 100 \times FracEL$.
- β_2 = efeito de aumentar 1 unidade de $PctEL$ mantendo constante $FracEL_i$
- Programas estatísticos geralmente eliminam automaticamente uma das variáveis

33 Armadilha da dummy

Duas variáveis indicadoras para sexo de nascimento:

$$H_i = \begin{cases} 1, & \text{se } i \text{ é homem,} \\ 0, & \text{se } i \text{ é mulher.} \end{cases}$$

$$M_i = \begin{cases} 1, & \text{se } i \text{ é mulher,} \\ 0, & \text{se } i \text{ é homem.} \end{cases}$$

...

$Y_i = \beta_0 + \beta_1 H_i + \beta_2 M_i + u_i$: não pode ser estimado!

- Multicolinearidade perfeita: $H_i + M_i = 1$

...

Pode-se estimar:

1. $Y_i = \beta_0 + \beta_1 H_i + u_i$
2. $Y_i = \beta_0 + \beta_2 M_i + u_i$
3. $Y_i = \beta_1 H_i + \beta_2 M_i + u_i$

34 Regra para evitar armadilha das dummies

- Se você tem K variáveis indicadoras em que cada observação cai em uma (e somente uma) categoria, não é possível estimar coeficientes para todas as K mais o intercepto

- Solução padrão: incluir na regressão $K - 1$ variáveis e o intercepto
 - neste caso, o coeficiente de uma variável indicadora = diferença entre aquela categoria e a categoria excluída
- Também é possível incluir todas as K variáveis e excluir o intercepto