FGV EPGE
ESCOLA BRASILEIRA DE ECONOMIA E FINANÇAS

Notas de Aula para o Mestrado e Doutorado em Economia

FGV/EPGE — Professor Marcelo Mello

Vitor Wilher[1]

Versão preliminar — compilada ao longo do curso (Agosto–Outubro de 2026). Baseada nos slides da disciplina (Prof. Marcelo Mello) e em Stock e Watson (2020).

[1]Bacharel e Mestre em Economia pela UFF, Candidato ao PhD em Economia pela EPGE/FGV. É Especialista em Ciências de Dados e Inteligência Artificial Generativa pela PUC-Rio. Atualmente, exerce a função de Data Tech Lead na Análise Macro (http://analisemacro.com.br). Saiba mais em https://github.com/vitorwilher.

Índice

1 Introdução

Estas notas acompanham a disciplina de **Econometria** do Mestrado e Doutorado Profissional em Economia e Finanças (MPEF) da FGV/EPGE, ministrada pelo Professor Marcelo Mello entre agosto e outubro de 2026. Elas seguem a ordem das aulas e tomam os slides da disciplina como fonte primária: a notação, os exemplos e o encadeamento são os do professor, ainda que o texto seja de minha autoria e acrescente derivações, comentários e implementações computacionais que os slides deixam implícitos.

> **!** Estado destas notas
>
> O curso tem **quinze aulas**. Esta versão cobre as **quatro primeiras**, que foram as ministradas até o momento da compilação: causalidade e estrutura de dados (Aula 1), revisão de probabilidade (Aula 2), o modelo de regressão linear com um regressor (Aula 3) e a inferência sobre seus coeficientes — testes de hipótese, intervalos de confiança, heterocedasticidade e Gauss-Markov (Aula 4). Os capítulos seguintes serão acrescentados conforme o curso avança — a estrutura do documento já está montada para recebê-los.

1.1 O que é econometria

A definição que abre o curso é deliberadamente ampla: *a econometria é a arte e a ciência de utilizar teoria econômica e técnicas estatísticas para analisar dados econômicos.* As duas palavras importam. É **ciência** porque há teoremas, estimadores com propriedades demonstráveis e procedimentos de inferência com garantias precisas. É **arte** porque nenhum teorema decide, no lugar do pesquisador, se a hipótese de identificação é plausível no problema concreto que ele tem diante de si — e é justamente aí que o trabalho empírico se ganha ou se perde.

O professor sintetiza a diferença de postura em um contraste: *economista* versus *estatístico*. O estatístico pergunta se a associação nos dados é real ou fruto do acaso amostral; o economista quer saber, além disso, se ela é **causal** — se intervir na variável X mudaria a variável Y. A segunda pergunta é bem mais difícil, e é a que organiza o curso inteiro.

1.2 Por que ela importa

Muitas decisões — de política pública, de empresa, de carreira — dependem de compreender as relações entre as variáveis do mundo. E essas decisões exigem **respostas quantitativas para perguntas quantitativas**: não basta saber que reduzir o tamanho da turma "ajuda"; é preciso saber *quanto* ajuda, para comparar com o custo de contratar professores.

Há ainda uma razão contemporânea. Em um mundo dominado por *big data*, em que a coleta de dados virou padrão e estudos empíricos se multiplicam no setor público e privado, a análise e a interpretação competentes tornaram-se instrumento essencial de decisão — e, igualmente importante, **instrumento de identificação de erros** nas análises alheias. Boa parte do valor de um econometrista treinado está em saber reconhecer quando um resultado empírico não sustenta a conclusão que anuncia. O Seção 2 abre com quatro histórias em que exatamente isso falhou, com consequências que se contam em dezenas de milhares de mortes.

1.3 O fio condutor

O curso tem uma tese, repetida do primeiro ao último slide: **correlação não implica causalidade**. O que dá conteúdo a essa frase — que de outro modo seria um clichê — é o programa de trabalho que ela impõe.

Começamos reconhecendo o problema (Seção 2): queremos efeitos causais, o padrão ouro para obtê-los é o experimento aleatorizado, e quase nunca temos um. Trabalhamos com **dados observacionais**, gerados por escolhas dos agentes, não por sorteio em laboratório. Disso decorrem as duas dificuldades que o curso passa o resto do tempo enfrentando: **variáveis omitidas** e **causalidade simultânea**.

Em seguida recuperamos as ferramentas probabilísticas necessárias (Seção 3) — esperança, variância, distribuição conjunta e, sobretudo, a **média condicional**, que é o objeto central de tudo o que vem depois. A regressão, afinal, é um método para estimar médias condicionais.

Então construímos o instrumento básico: o **modelo de regressão linear com um regressor** e o estimador de **mínimos quadrados ordinários** (1). Aqui aparece a hipótese que carrega todo o peso da inferência causal, $\mathbb{E}(u_i \mid X_i) = 0$, e a constatação desconfortável de que ela **não é testável** — sua plausibilidade depende de conhecimento do problema e de julgamento, não de estatística.

De posse do estimador, aprendemos a medir a incerteza que ele carrega (Seção 5): testes de hipótese, intervalos de confiança e a leitura completa de uma tabela de regressão. Aqui aparecem também dois resultados que organizam o que vem depois — o de que a regressão sobre uma *dummy* é literalmente uma **comparação de médias**, o que reconcilia o instrumento com o arcabouço experimental do Seção 2; e o de que a **heterocedasticidade** compromete a inferência sem viesar as estimativas, razão pela qual os erros-padrão robustos são o padrão da prática aplicada.

O restante do curso, ainda por vir, é uma sequência de estratégias para tornar essa hipótese defensável: controlar por variáveis observáveis (regressão múltipla), explorar variação temporal (dados em painel, efeitos fixos), encontrar variação exógena (variáveis instrumentais) ou aproximar experimentos com quase-experimentos (diferenças-em-diferenças, regressão descontínua, controle sintético, *propensity score matching*).

1.4 Como ler estas notas

Cada capítulo abre com a **motivação** do problema, desenvolve a teoria com **derivações passo a passo** e fecha com duas seções fixas:

- um **Laboratório computacional em R**, que reproduz numericamente os resultados do capítulo ou gera a figura que o ilustra. O código roda na própria compilação do PDF, de modo que os números impressos são os que o R de fato produziu;
- uma lista de **Exercícios**, no espírito das atividades da disciplina.

Resultados centrais aparecem em caixa, assim. Hipóteses que não podem ser esquecidas vão em caixas de destaque. Quando um conceito de estatística é mobilizado — esperança condicional, distribuição amostral, teorema central do limite —, uma caixa *Ferramenta estatística* o nomeia e remete ao capítulo correspondente das **Notas de Aula de Estatística** (Prof. Eduardo Campos), escritas para o mesmo programa.

Uma observação sobre software. Os slides do curso mostram saídas do **STATA**, que é o padrão na disciplina. Os laboratórios destas notas usam **R**, por ser o que utilizo no trabalho, e reproduzem as

mesmas quantidades — inclusive os **erros-padrão robustos**, que implemento a partir da fórmula para deixar explícito o que o comando `robust` do STATA faz por baixo do capô.

1.5 Referências

O texto base do curso é Stock e Watson (2020), e a numeração dos capítulos dos slides o segue de perto. Como auxiliares, o programa lista Wooldridge (2016), Angrist e Pischke (2015) e Angrist e Pischke (2009) — os dois últimos são a referência natural para a parte de inferência causal e desenho de pesquisa que fecha o curso. Onde estas notas avançam além do slide, indico a fonte no próprio texto.

2 Econometria e Causalidade

O curso não começa por definições nem por fórmulas. Começa por **quatro histórias** em que a análise empírica foi feita, publicada em periódicos de primeira linha, aceita por reguladores — e estava errada. Custaram, em um dos casos, algo entre 26 e 38 mil vidas.

A escolha é deliberada. É fácil aprender a rodar uma regressão; o difícil é adquirir o julgamento para saber quando uma regressão — a sua ou a de outra pessoa — não sustenta a conclusão que anuncia. As histórias existem para instalar esse desconforto produtivo desde a primeira aula.

2.1 Quatro histórias

2.1.1 Cigarro e câncer: quando a associação estava certa

O primeiro caso é o contraponto dos demais. Em 1950, Wynder e Graham (1950) publicaram no *JAMA* um estudo com 684 pacientes de câncer de pulmão que encontrou forte associação estatística com o consumo pesado de cigarros. Em 1954, Doll e Hill (1954) acompanharam médicos britânicos ao longo do tempo e mapearam a mortalidade crescente contra o consumo de tabaco.

Hoje ninguém duvida da conclusão. Mas vale registrar que, à época, ela foi contestada com um argumento **estatisticamente impecável**: nenhum dos dois estudos era um experimento aleatorizado, de modo que a associação poderia refletir uma terceira variável — uma predisposição genética que causasse, simultaneamente, o hábito de fumar e o câncer. Ronald Fisher, um dos fundadores da estatística moderna, defendeu exatamente essa posição.

A objeção era logicamente válida e, ainda assim, a conclusão substantiva era correta. Fica a primeira lição: **a crítica de variável omitida é sempre disponível, e por isso mesmo precisa ser avaliada, não apenas invocada**. O que eventualmente resolveu a questão não foi um teorema, foi o acúmulo de evidência de desenhos diferentes — dose-resposta, mecanismo biológico, coortes, redução de risco após cessação.

2.1.2 Vioxx: o experimento perfeito que escondeu o essencial

O segundo caso é o mais grave, e o mais instrutivo, porque **aqui houve experimento aleatorizado** — e ele não bastou.

A Merck, eleita "empresa mais admirada" pela *Fortune* entre 1985 e 1990, desenvolveu o rofecoxib (Vioxx) com uma ideia farmacológica elegante. Os anti-inflamatórios não esteroides tradicionais — aspirina, ibuprofeno, naproxeno — inibem a enzima COX-2, que causa inflamação e dor, mas também inibem a COX-1, que protege o revestimento do estômago; daí as úlceras. O Vioxx foi desenhado para **inibir a COX-2 sem inibir a COX-1**, e foi aprovado pela FDA após "rigorosos ensaios clínicos aleatorizados".

O estudo VIGOR (Bombardier et al. 2000) alocou aleatoriamente 8.076 pacientes com artrite reumatoide para receber 50mg diários de rofecoxib ou 500mg de naproxeno duas vezes ao dia. O desfecho primário eram eventos gastrointestinais superiores confirmados. O resultado confirmou a hipótese: 2,1 eventos por 100 pacientes-ano com rofecoxib contra 4,5 com naproxeno. A conclusão publicada foi que o rofecoxib se associa a "significativamente menos eventos gastrointestinais clinicamente importantes".

O que deu errado? Duas coisas, e nenhuma delas é falha de aleatorização.

Primeiro, **os tomadores de Vioxx tiveram 17 infartos contra 4 dos tomadores de naproxeno — e a diferença foi interpretada como estatisticamente não significante**. Segundo, uma informação foi omitida do relato: 47 pacientes em uso de Vioxx tiveram eventos tromboembólicos, contra 20 no grupo naproxeno.

A Merck retirou o Vioxx do mercado em 30 de setembro de 2004. As estimativas de dano vão de 88 a 139 mil infartos, com 26 a 38 mil mortes.

> **!** O que o caso Vioxx ensina
>
> A aleatorização garante que os grupos sejam comparáveis — e ela funcionou aqui. O que ela **não** garante é que o desfecho certo esteja sendo medido, que a amostra tenha poder suficiente para detectar o efeito adverso, ou que todos os resultados sejam reportados.
> "Não significante" **não** é sinônimo de "não existe": com 17 contra 4 eventos, o intervalo de confiança era largo demais para excluir um efeito grande. Ausência de evidência não é evidência de ausência — uma distinção que retomaremos quando estudarmos poder de teste.

2.1.3 Ovos perigosos: uma hipótese plausível que virou política

Por décadas, autoridades médicas sustentaram que comer ovos ameaçava a saúde cardiovascular. A origem é a **hipótese lipídica** (1950–1970), que ligava colesterol sanguíneo elevado a maior risco cardíaco. A lógica parecia inescapável: a gema é excepcionalmente rica em colesterol dietético, logo comer ovos elevaria o colesterol sanguíneo e obstruiria artérias.

Em 1968 a American Heart Association recomendou oficialmente não mais que 300mg diários de colesterol dietético e restringiu o consumo a três ovos por semana. A recomendação sobreviveu por quase cinco décadas antes de ser revista.

O erro aqui não foi estatístico, foi de **inferência sobre o mecanismo**: assumiu-se que colesterol ingerido se traduz em colesterol circulante, quando a regulação hepática compensa boa parte da ingestão. Uma cadeia causal plausível, adotada sem o teste da etapa intermediária.

2.1.4 Cloroquina: dose, mortalidade e o valor de um controle

O quarto caso é recente e brasileiro. Em abril de 2020, um estudo conduzido em Manaus com 81 pacientes graves de covid-19 (Borba et al. 2020) comparou duas doses de cloroquina: um grupo recebeu dose alta (600mg duas vezes ao dia por dez dias), outro recebeu dose baixa (450mg por cinco dias, duas vezes apenas no primeiro dia). O grupo de dose alta apresentou letalidade **17% maior**.

O que torna este estudo informativo, apesar do tamanho pequeno, é justamente ter havido **comparação entre grupos alocados**. Sem o braço de dose baixa, a mortalidade observada no grupo de dose alta seria atribuída à gravidade da doença — e a droga poderia parecer, na melhor das hipóteses, inócua.

2.2 O que é econometria

Recolhidas as histórias, a definição do curso:

> **!** Definição
>
> De forma ampla, a **econometria** é a arte e a ciência de utilizar teoria econômica e técnicas estatísticas para analisar dados econômicos.

O professor contrasta duas figuras: o **economista** e o **estatístico**. O estatístico domina as técnicas de descrever associações e quantificar incerteza amostral. O economista traz a pergunta substantiva e a teoria que disciplina a interpretação — em particular, a distinção entre associação e efeito causal. A econometria vive na interseção, e nenhuma das duas metades é dispensável.

Por que estudá-la? Muitas decisões dependem de compreender as relações entre as variáveis do mundo, e essas decisões exigem **respostas quantitativas para perguntas quantitativas**. Além disso, em um mundo dominado por *big data*, em que a coleta de dados virou padrão e estudos empíricos proliferam no setor público e privado, a análise competente tornou-se instrumento essencial de decisão — e, tão importante quanto, **instrumento para identificar erros nas análises empíricas alheias**. As quatro histórias acima são o argumento.

2.3 Duas questões econômicas

O curso se organiza em torno de perguntas concretas. Duas delas atravessam as primeiras aulas.

Questão 1: reduzir o tamanho da turma melhora o desempenho acadêmico? É a pergunta que gera o exemplo `TestScore/ClassSize` do 1 e que reaparece no curso inteiro. Tem interesse imediato de política pública: contratar professores é caro, e o tradeoff só se avalia com uma estimativa quantitativa do efeito.

Questão 2: há discriminação salarial por gênero? O caso *Boylan v. New York Times* mostra a econometria em um tribunal. Em novembro de 1974, após a formação de um *Women's Caucus* em 1972, seis mulheres — Elizabeth Boylan, Louise Carini, Joan Cook, Nancy Davis, Grace Glueck e Andrea Skinner — abriram ação federal representando cerca de 600 funcionárias, sob o Título VII do Civil Rights Act de 1964. As alegações eram viés sistemático em remuneração, atribuição desigual de funções e barreiras de promoção.

A estratégia probatória foi estatística. Sob coordenação de Bruce Levin, da Universidade Columbia, as autoras apresentaram uma **regressão linear múltipla** sobre registros de aproximadamente 1.800 empregados sindicalizados, com um conjunto de **99 variáveis explicativas** controlando fatores de produtividade — escolaridade, anos de experiência, tempo de casa. O achado central: mesmo após controlar pelas 99 variáveis, **o coeficiente de gênero permaneceu negativo e estatisticamente significante**.

O argumento tem a estrutura exata da inferência causal com dados observacionais: se todos os determinantes legítimos do salário foram controlados, o que resta associado ao gênero não pode ser explicado por produtividade. Note que a força do argumento depende inteiramente da **completude do conjunto de controles** — é precisamente o ponto que a hipótese $\mathbb{E}(u \mid X) = 0$ formaliza no 1, e é onde a defesa concentraria seu contra-argumento.

2.4 O que significa causalidade

> **!** Definição
>
> **Causalidade** significa que uma ação específica leva a uma consequência mensurável específica. De forma simples, se uma mudança em X causa Y, então uma variação em X produz variação em Y, **tudo mais constante**.

A cláusula *ceteris paribus* é onde mora toda a dificuldade. No mundo, as coisas não variam uma de cada vez.

2.4.1 O experimento aleatorizado ideal

A forma canônica de estimar efeitos causais é o **experimento controlado aleatorizado ideal**. Ele é *controlado* no sentido de que há um **grupo de tratamento**, que recebe o tratamento, e um **grupo de controle**, que não recebe. E é *aleatorizado* no sentido de que a alocação ao tratamento é feita **por sorteio**.

Nesse arcabouço, o **efeito causal** é definido como o efeito de uma ação ou tratamento sobre um resultado, medido em um experimento aleatorizado ideal. A propriedade decisiva é esta:

> Nesse experimento, a única razão sistemática para as diferenças de resultado entre os grupos de tratamento e controle é o próprio tratamento.

A aleatorização faz com que os dois grupos sejam, **em média**, idênticos em tudo — idade, renda, motivação, saúde prévia, e também em características que ninguém mediu ou sequer imaginou medir. É essa equivalência em expectativa que autoriza atribuir a diferença de resultados ao tratamento. Formalmente, é o que garante que a diferença de médias

$$\mathbb{E}(Y \mid X = 1) - \mathbb{E}(Y \mid X = 0)$$

seja um estimador não viesado do efeito causal — resultado que o 1 retoma ao apresentar as hipóteses de MQO.

2.5 O problema fundamental

Aqui o curso enuncia sua dificuldade central:

> **!** O problema que organiza o curso
>
> Desejamos obter uma estimativa do **efeito causal**. Mas trabalhamos com **dados observacionais**, o que elimina a possibilidade de alocação aleatória do tratamento.
> Ou seja: sabemos que **houve escolha**, e que existem **variáveis omitidas**.

Dados observacionais são gerados a partir de escolhas das partes envolvidas — quem estuda mais, quem toma o remédio, qual escola adota turmas menores —, não por um processo aleatório em

laboratório. E quem escolhe, escolhe por razões, que em geral também afetam o resultado de interesse.

A resposta da econometria de regressão é uma aposta explícita:

> A inferência causal baseada em um modelo de regressão tem como pilar fundamental a hipótese de que, uma vez que se controla os grupos de tratamento e controle pelas **variáveis-chave observáveis**, o viés de seleção causado por variáveis omitidas também é eliminado.

É exatamente o que os autores do caso *New York Times* tentaram fazer com 99 controles. E é uma **hipótese**, não um teorema — o que os casos da cloroquina e dos ovos ilustram pelo avesso.

2.5.1 As quatro perguntas do projeto ideal

Angrist e Pischke (Angrist e Pischke 2009) organizam o desenho de pesquisa em quatro perguntas, que o professor adota como roteiro:

1. Qual é a **relação causal** de interesse?
2. Qual é o **experimento** que idealmente captaria esse efeito causal?
3. Qual é a sua **estratégia de identificação**?
4. Qual é o seu **método de inferência estatística**?

A segunda pergunta é a mais subestimada. Mesmo quando o experimento é inviável — por custo, por ética, por impossibilidade lógica —, imaginá-lo com precisão disciplina todo o resto: ele define o que seria o contrafactual, e portanto o que a estratégia observacional precisa aproximar.

> **!** Estratégia de identificação
>
> A **estratégia de identificação** é a maneira pela qual um pesquisador usa dados observacionais para **aproximar um experimento real**.

Boa parte do curso — variáveis instrumentais, diferenças-em-diferenças, regressão descontínua, controle sintético — é um catálogo de estratégias de identificação.

2.5.2 Perguntas que não cabem no arcabouço

Nem toda pergunta admite a moldura tratamento-controle. Questões sobre o efeito causal de **raça ou gênero** são difíceis de colocar nesses termos, porque essas características raramente podem ser manipuladas isoladamente: não há como sortear o gênero de uma pessoa mantendo tudo o mais constante, e mesmo conceitualmente é obscuro o que significaria "a mesma pessoa, de outro gênero".

Isso não torna a pergunta ilegítima — o caso *New York Times* é prova de que ela é socialmente urgente. Torna-a uma pergunta cuja resposta empírica exige cuidado redobrado com o que exatamente está sendo estimado. É um tema que o curso retoma adiante.

2.6 Estrutura de dados

Fechada a discussão conceitual, a taxonomia operacional. Em econometria, os dados vêm de duas fontes: **experimentais**, gerados em laboratório ou por algum processo aleatório deliberado; e **não experimentais**, isto é, **observacionais**.

Quanto à organização, há três estruturas. Sejam i o índice da **entidade** (indivíduo, firma, país) e t o índice do **tempo calendário**:

Estrutura	Índices	Exemplo
Corte transversal (*cross-section*)	$i = 1, \dots, n$ com t fixo	Efeito da educação no salário para 5.000 indivíduos em 2017
Série de tempo	i fixo, $t = 1, \dots, T$	Efeito dinâmico de um choque no petróleo sobre a inflação
Painel (ou longitudinal)	$i = 1, \dots, n$ **e** $t = 1, \dots, T$	Determinantes da produtividade de países latino-americanos ao longo do tempo

O painel combina as duas dimensões, e é essa combinação que o torna valioso: observar a mesma entidade em momentos diferentes permite controlar por tudo aquilo que é fixo na entidade e não observado — a base do estimador de efeitos fixos, que o curso verá adiante.

Uma convenção de nomenclatura vale registro. Painéis com n **grande** e T **pequeno** são típicos de problemas de **microeconomia** — milhares de indivíduos, poucos anos. Painéis com n pequeno e T grande são típicos de **macroeconomia** — poucos países, muitas décadas. A distinção não é cosmética: ela determina em que dimensão a teoria assintótica opera e, portanto, quais estimadores e erros-padrão são apropriados.

Este curso trata de **corte transversal e painel**; séries de tempo ficam para outra disciplina.

2.7 As duas dificuldades

A maior parte do curso trata das dificuldades associadas ao uso de dados observacionais para estimar efeitos causais. Elas são duas:

- **Variáveis omitidas** (*confounding*): um fator não incluído no modelo afeta simultaneamente o tratamento e o resultado. É o caso da predisposição genética na controvérsia do cigarro, e o que os 99 controles do caso *New York Times* tentavam neutralizar;
- **Causalidade simultânea**: X afeta Y, mas Y também afeta X. Turmas menores podem melhorar o desempenho, mas escolas com desempenho ruim podem receber verba adicional para reduzir turmas — e aí a correlação medida mistura os dois sentidos.

Ambas serão formalizadas no 1 como violações da hipótese $\mathbb{E}(u_i \mid X_i) = 0$, e é para contorná-las que existem as estratégias de identificação da segunda metade do curso.

> **!** A frase que o curso repetirá muitas vezes
>
> É **imediato** estabelecer correlação. O problema todo é que **correlação não implica causalidade**.

2.8 Laboratório computacional em R

O laboratório deste capítulo é conceitual: quero mostrar, com números, por que a aleatorização funciona e o que exatamente se perde sem ela.

2.8.1 Por que a aleatorização funciona

Simulo uma população em que o tratamento tem efeito causal **conhecido** de +5 unidades, e em que uma variável de confusão — chame-a *motivação* — afeta tanto a probabilidade de receber tratamento quanto o resultado. É a estrutura do problema de seleção.

```
set.seed(2026)
n <- 20000
motivacao <- rnorm(n)                       # confundidor NAO observado
efeito_causal <- 5                          # o parametro que queremos recuperar

# resultado potencial sem tratamento: depende da motivacao
Y0 <- 50 + 8 * motivacao + rnorm(n, sd = 5)
Y1 <- Y0 + efeito_causal                    # efeito constante, por simplicidade

# ESCOLHA: mais motivados se autosselecionam ao tratamento
p_trat  <- plogis(1.5 * motivacao)
D_obs   <- rbinom(n, 1, p_trat)             # tratamento observacional
D_ale   <- rbinom(n, 1, 0.5)                # tratamento aleatorizado

c(efeito_verdadeiro = efeito_causal,
  prop_tratada_obs  = mean(D_obs),
  prop_tratada_ale  = mean(D_ale))
```

```
efeito_verdadeiro  prop_tratada_obs  prop_tratada_ale
          5.00000           0.50210           0.50005
```

Com os dados montados, comparo as duas estimativas ingênuas de diferença de médias:

```
Y_obs <- ifelse(D_obs == 1, Y1, Y0)   # o que se observa no mundo observacional
Y_ale <- ifelse(D_ale == 1, Y1, Y0)   # o que se observaria com sorteio

dif_obs <- mean(Y_obs[D_obs == 1]) - mean(Y_obs[D_obs == 0])
dif_ale <- mean(Y_ale[D_ale == 1]) - mean(Y_ale[D_ale == 0])
```

```r
c(verdadeiro = efeito_causal,
  observacional = dif_obs, vies_observacional = dif_obs - efeito_causal,
  aleatorizado  = dif_ale, vies_aleatorizado  = dif_ale - efeito_causal)
```

```
       verdadeiro      observacional vies_observacional       aleatorizado
        5.00000000       13.53237861         8.53237861         4.98688525
vies_aleatorizado
       -0.01311475
```

O contraste é o capítulo inteiro em três linhas. A comparação **observacional** superestima grosseiramente o efeito: ela atribui ao tratamento não só os 5 pontos que ele de fato causa, mas também a vantagem prévia dos mais motivados, que se autosselecionaram. A comparação **aleatorizada** acerta o alvo, com erro compatível com ruído amostral.

A razão aparece ao inspecionar o **balanceamento** do confundidor entre os grupos:

```r
rbind(
  observacional = c(tratados = mean(motivacao[D_obs == 1]),
                    controles = mean(motivacao[D_obs == 0]),
                    diferenca = mean(motivacao[D_obs == 1]) -
                                mean(motivacao[D_obs == 0])),
  aleatorizado  = c(mean(motivacao[D_ale == 1]),
                    mean(motivacao[D_ale == 0]),
                    mean(motivacao[D_ale == 1]) - mean(motivacao[D_ale == 0]))
)
```

```
               tratados   controles   diferenca
observacional 0.53770839 -0.52101584 1.058724233
aleatorizado  0.01545656  0.00568165 0.009774915
```

Sob aleatorização, os grupos têm motivação média praticamente idêntica — a diferença é ruído. Sob autosseleção, os tratados são sistematicamente mais motivados, e é essa diferença prévia que contamina a comparação de resultados. **A aleatorização não elimina o confundidor; ela o distribui igualmente entre os grupos**, que é tudo de que precisamos.

```r
rot_obs <- "observacional (autosselecao)"
rot_ale <- "aleatorizado (sorteio)"
d <- rbind(
  data.frame(motivacao, grupo = ifelse(D_obs == 1, "tratado", "controle"),
             desenho = rot_obs),
  data.frame(motivacao, grupo = ifelse(D_ale == 1, "tratado", "controle"),
             desenho = rot_ale)
)
# fixa a ordem dos paineis: autosselecao a esquerda, sorteio a direita
d$desenho <- factor(d$desenho, levels = c(rot_obs, rot_ale))
```

```r
ggplot(d, aes(x = motivacao, fill = grupo)) +
  geom_density(alpha = 0.45, colour = NA) +
  facet_wrap(~desenho) +
  scale_fill_manual(values = c(controle = cz, tratado = az)) +
  labs(x = "motivacao (confundidor nao observado)", y = "densidade")
```

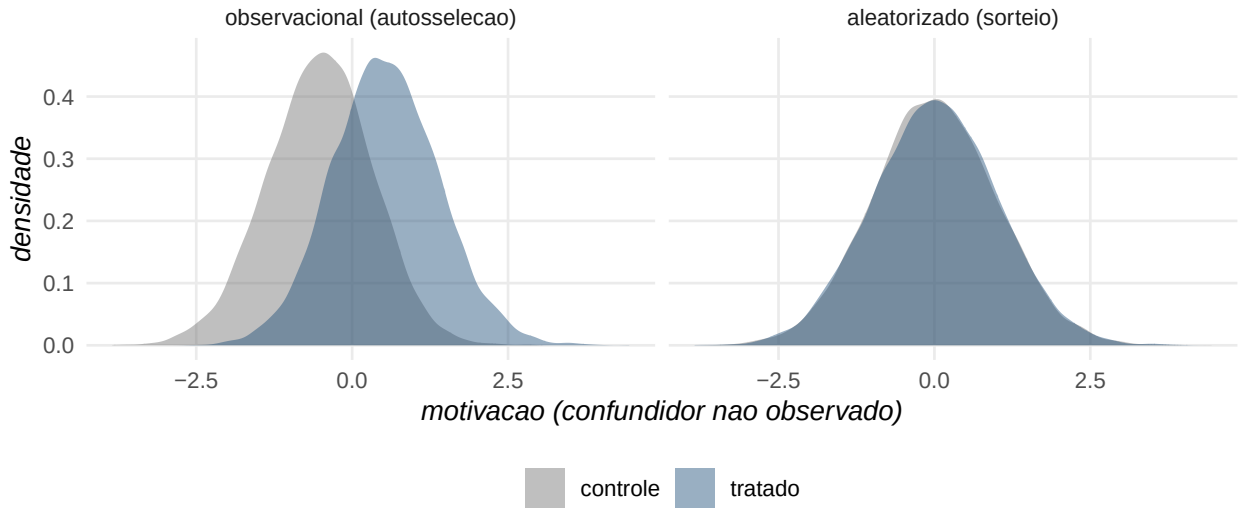

Figura 1: Distribuição do confundidor (motivação) entre tratados e controles. Sob autosseleção (esquerda) os grupos são visivelmente diferentes antes de qualquer tratamento; sob aleatorização (direita) as distribuições se sobrepõem.

2.8.2 O que "controlar por observáveis" resolve — e o que não resolve

A aposta da regressão é que, controlando pelos observáveis, o viés desaparece. Vale ver isso funcionar e, em seguida, falhar.

```r
# caso 1: o confundidor E observado -> controlar resolve
m_sem  <- lm(Y_obs ~ D_obs)
m_com  <- lm(Y_obs ~ D_obs + motivacao)

# caso 2: observamos apenas um PROXY ruidoso da motivacao
proxy <- motivacao + rnorm(n, sd = 1.2)
m_proxy <- lm(Y_obs ~ D_obs + proxy)

round(c(verdadeiro         = efeito_causal,
        sem_controle       = coef(m_sem)[2],
        com_confundidor    = coef(m_com)[2],
        com_proxy_ruidoso  = coef(m_proxy)[2]), 4)
```

```
            verdadeiro      sem_controle.D_obs   com_confundidor.D_obs
```

```
                5.0000                  13.5324                  5.0189
com_proxy_ruidoso.D_obs
                10.6181
```

Três resultados, três lições. Sem controle, o viés é grande. Com o confundidor **observado e incluído**, a regressão recupera o efeito causal quase exatamente — é a aposta da inferência causal por regressão funcionando como prometido. Com um **proxy ruidoso** do confundidor, o viés diminui bastante mas **não desaparece**: controlar imperfeitamente corrige imperfeitamente.

O terceiro caso é o realista. Raramente observamos o confundidor; em geral temos indicadores imperfeitos dele. Daí a advertência do curso de que a hipótese de identificação é uma **aposta substantiva**, que depende de conhecimento do problema — e não algo que os dados possam confirmar sozinhos.

2.9 Exercícios

1. **A crítica de Fisher.** Fisher argumentou que a associação entre fumo e câncer poderia ser explicada por um gene que causasse ambos. (a) Escreva essa hipótese como um problema de variável omitida. (b) Que tipo de evidência empírica poderia distinguir a hipótese causal da genética? (c) Por que o argumento de Fisher, sendo logicamente válido, não é suficiente para descartar a conclusão substantiva?

2. **Vioxx e poder de teste.** No estudo VIGOR observaram-se 17 infartos no grupo rofecoxib contra 4 no naproxeno, e a diferença foi julgada não significante. (a) Calcule a razão de incidência. (b) Discuta por que "não significante" e "não existe efeito" são afirmações distintas. (c) Que informação, além do valor-p, deveria ter sido reportada?

3. **Desenhando o experimento ideal.** Para cada pergunta abaixo, descreva o experimento aleatorizado ideal — mesmo que inviável — e explique o que o torna inviável: (a) o efeito de um ano adicional de escolaridade sobre o salário; (b) o efeito do salário mínimo sobre o emprego; (c) o efeito de uma fusão empresarial sobre os preços.

4. **As quatro perguntas.** Escolha um artigo empírico de economia que você conheça e responda, para ele, as quatro perguntas de Angrist e Pischke. Onde está o elo mais fraco do argumento?

5. **Causalidade simultânea.** Suponha que turmas menores melhorem o desempenho, mas que escolas com desempenho ruim recebam verba para reduzir turmas. (a) Qual o sinal esperado do viés na regressão de desempenho contra tamanho de turma? (b) A correlação medida subestima ou superestima o efeito causal? Justifique.

6. **O caso *New York Times*.** A defesa poderia argumentar que faltou uma variável ao conjunto de 99 controles. (a) Que variável seria essa, e por que ela precisaria afetar simultaneamente gênero e salário? (b) Por outro lado, incluir controles em excesso também pode ser problemático — o que aconteceria se um dos 99 controles fosse *ele próprio* consequência da discriminação (por exemplo, o cargo ocupado)?

7. **Estruturas de dados.** Classifique como corte transversal, série de tempo ou painel:

(a) PIB trimestral brasileiro de 1996 a 2025; (b) a PNAD Contínua de um trimestre específico; (c) as demonstrações financeiras anuais de 200 empresas ao longo de 10 anos. Para (c), que tipo de variável omitida o painel permitiria controlar que o corte transversal não permitiria?

8. **Simulação.** Modifique o laboratório deste capítulo para que o efeito do tratamento seja **heterogêneo** — por exemplo, maior para os mais motivados. (a) A diferença de médias sob aleatorização ainda recupera algum parâmetro interpretável? Qual? (b) E a comparação observacional? (c) Que implicação isso tem para a leitura de resultados experimentais quando a população é heterogênea?

3 Revisão de Probabilidade

O Seção 2 terminou com um problema, não com uma solução: queremos efeitos causais, o padrão ouro para obtê-los é o experimento aleatorizado, e quase nunca temos um. Antes de construir a primeira ferramenta capaz de enfrentar esse problema — o modelo de regressão linear —, precisamos recuperar o vocabulário probabilístico em que ela é escrita. É o que faz este capítulo.

A revisão não é um preâmbulo protocolar. Há **duas razões precisas** para um curso de econometria começar por probabilidade, e vale enunciá-las antes de qualquer definição, porque elas organizam tudo o que vem a seguir.

A primeira é que **a regressão é, essencialmente, um método para estimar médias condicionais**. Quando escrevemos $Y_i = \beta_0 + \beta_1 X_i + u_i$ e impomos $\mathbb{E}(u_i \mid X_i) = 0$, o que estamos de fato afirmando é que a reta populacional $\beta_0 + \beta_1 X$ **é** a média de Y dado X. Toda a interpretação de β_1 como "efeito de X sobre Y" passa por aí. Sem entender o que é a média condicional $\mathbb{E}(Y \mid X)$ — o que ela mede, como se comporta, em que sentido ela difere de correlação —, o modelo de regressão vira manipulação simbólica sem conteúdo. Por isso a **média condicional** e a **lei das expectativas iteradas** recebem, aqui, tratamento desproporcional ao seu tamanho nos slides: são as peças que reaparecerão, sem aviso, em cada passo de 1.

A segunda razão é a inferência. Um estimador é uma função dos dados; os dados são aleatórios; logo o estimador é uma **variável aleatória**, com distribuição própria. Tudo o que dizemos sobre precisão — erros-padrão, intervalos de confiança, testes de hipótese — é uma afirmação sobre essa **distribuição amostral**. Esperança e variância, que revisamos adiante, são exatamente os objetos com que se calcula viés e precisão de um estimador.

> **!** Convenção de notação
>
> Usamos letras maiúsculas (X, Y) para variáveis aleatórias e minúsculas (x, y) para os valores que elas podem assumir. Parâmetros populacionais são letras gregas: μ para médias, σ^2 para variâncias, θ para a probabilidade de sucesso de uma Bernoulli. Todo este capítulo trata de **quantidades populacionais** — não de estimativas amostrais. A passagem da população para a amostra é o assunto de 1.

3.1 Variáveis aleatórias e distribuição de probabilidade

Uma **variável aleatória** é uma variável cujo valor numérico depende de um experimento aleatório. O exemplo mínimo do professor é o lançamento de uma moeda: o experimento tem dois resultados possíveis, cara e coroa, e a variável aleatória é a regra que associa um número a cada um deles.

Defina X assumindo o valor 1 se sai cara — evento com probabilidade θ — e zero caso contrário:

$$X = \begin{cases} 1, & \text{com probabilidade } \theta, \\ 0, & \text{com probabilidade } 1 - \theta. \end{cases}$$

Essa é a **variável aleatória de Bernoulli**. Apesar da simplicidade, ela é onipresente em economia aplicada: participar ou não do mercado de trabalho, receber ou não um tratamento, uma empresa

entrar em default ou não. Boa parte das variáveis dependentes interessantes em microeconometria são Bernoullis.

Variáveis aleatórias são **discretas** ou **contínuas**. A Bernoulli é discreta: assume valores em um conjunto discreto, como $0, 1, 2, ...$ Uma variável aleatória **contínua** assume valores em um *continuum* — tipicamente resultados de medições (peso, tempo, renda, retorno de um ativo) e não de contagens.

> **!** Definição — distribuição de probabilidade (caso discreto)
>
> A **distribuição de probabilidade** de uma variável aleatória discreta é a lista de todos os seus valores possíveis e as respectivas probabilidades associadas a cada valor.

3.1.1 A disputa de pênaltis e a Binomial

O professor constrói a distribuição Binomial a partir de um exemplo que vale a pena seguir com cuidado, porque o passo decisivo é justamente o que a intuição erra.

Considere uma disputa de pênaltis com **cinco cobranças**. O número de "sucessos" — isto é, pênaltis convertidos — é uma variável aleatória X que pode assumir os valores $x = 0, 1, 2, 3, 4, 5$. Cada cobrança é, isoladamente, uma Bernoulli(θ), com θ a probabilidade de converter, e as cobranças são independentes entre si.

Comece pelo caso fácil. A probabilidade de converter **todas** as cinco é

$$P(X = 5) = \theta^5,$$

porque há uma única forma de isso acontecer — a sequência $(1, 1, 1, 1, 1)$ — e a independência permite multiplicar as probabilidades.

Agora o caso interessante. Uma sequência com três sucessos e dois fracassos pode ocorrer, por exemplo, como

$$(1, 1, 1, 0, 0),$$

e **essa sequência em particular** ocorre com probabilidade $\theta^3(1-\theta)^2$. Aqui vem a pergunta do professor: essa **não** é a probabilidade de $P(X = 3)$. Por quê?

Porque existem várias formas de converter três gols. As sequências

$$(0, 1, 0, 1, 1), \qquad (0, 0, 1, 1, 1), \qquad (1, 0, 1, 0, 1), \qquad ...$$

são todas distintas como sequências, mas todas correspondem ao **mesmo** evento $X = 3$. Cada uma delas tem a mesma probabilidade $\theta^3(1-\theta)^2$ — a ordem não altera o produto —, e como são mutuamente excludentes, suas probabilidades se somam. Resta contar quantas são: é o número de formas de escolher quais 3 das 5 cobranças foram convertidas, ou seja, $C_3^5 = \binom{5}{3} = 10$. Portanto

$$P(X = 3) = C_3^5\, \theta^3(1-\theta)^2 = \frac{5!}{3!\,(5-3)!}\, \theta^3(1-\theta)^2.$$

O raciocínio não depende dos números escolhidos. De forma geral, para n tentativas independentes com probabilidade θ de sucesso em cada uma,

$$\boxed{P(X = x) = C_x^n\, \theta^x (1-\theta)^{\,n-x}, \qquad x = 0, 1, 2, \ldots, n}$$

que é a distribuição de probabilidade de uma **Binomial** com parâmetros (n, θ). Note a anatomia da fórmula: $\theta^x(1-\theta)^{n-x}$ é a probabilidade de *uma* sequência com x sucessos, e C_x^n conta *quantas* sequências produzem o mesmo total. O erro de omitir o coeficiente binomial é, no fundo, o erro de confundir uma sequência específica com um evento agregado.

> **i** Ferramenta estatística — análise combinatória e a Binomial
>
> A contagem de C_x^n sequências, a distinção entre arranjos e combinações e a dedução completa da família Binomial são desenvolvidas com mais vagar no capítulo de *Probabilidade e Distribuições Discretas* das **Notas de Aula de Estatística** (Prof. Eduardo Campos). Se o coeficiente binomial soa mecânico, vale a leitura: ele é a única parte da fórmula que não sai da independência.

3.1.2 Propriedades e distribuição acumulada

Seja X uma variável aleatória discreta que pode assumir n valores $x_1, x_2, \ldots, x_n$. Então valem duas propriedades que qualquer distribuição de probabilidade precisa satisfazer:

$$P(X = x_i) \geq 0 \quad \forall\, x_i,\ i = 1, 2, \ldots, n \qquad \text{e} \qquad \sum_{i=1}^{n} P(X = x_i) = 1.$$

A primeira é a exigência de que probabilidades não sejam negativas; a segunda, de que a lista de valores possíveis seja exaustiva — alguma coisa tem de acontecer.

A **distribuição de probabilidade acumulada** (ou função de distribuição acumulada, CDF) é a probabilidade de a variável aleatória ser menor ou igual a um valor particular. Para $X \sim \text{Bin}(n, \theta)$,

$$F(x_k) = P(X \leq x_k) = \sum_{i=1}^{k} P(X = x_i).$$

Voltando aos pênaltis com $n = 5$, a probabilidade de o time converter no máximo três cobranças é

$$P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3).$$

Duas observações sobre a CDF, ambas do professor e ambas consequência direta de as probabilidades serem não negativas e somarem um. Primeiro, a acumulada **nunca decresce**: cada termo adicionado à soma é ≥ 0, de modo que F é uma função não decrescente de x. Segundo, $P(X \leq x_n) = 1$ — ao chegar ao maior valor possível, a acumulada esgota toda a probabilidade.

3.2 Variáveis aleatórias contínuas

No caso contínuo, a probabilidade de ocorrer um valor específico é **zero**. O exemplo do professor é bom porque desarma a estranheza: qual é a probabilidade de alguém pesar exatamente 79,314159163 kg? Nenhuma balança do mundo distingue esse valor de seus vizinhos, e há um *continuum* não enumerável deles. A probabilidade de qualquer ponto isolado é nula.

O que faz sentido perguntar é a probabilidade de a variável cair em um **intervalo** — por exemplo, qual a probabilidade de o peso do indivíduo estar entre 75 kg e 85 kg. Formalmente,

$$P(a < X < b) = \int_a^b f(x)\,dx,$$

em que $f(\cdot)$ é a **função de densidade de probabilidade** (PDF). Como $P(X = a) = P(X = b) = 0$, incluir ou não os extremos é indiferente:

$$P(a < X < b) = P(a \le X \le b).$$

Essa é uma diferença qualitativa em relação ao caso discreto, onde trocar $<$ por $\le$ muda o resultado. A densidade satisfaz, em paralelo às duas propriedades do caso discreto,

$$f(x) \ge 0 \qquad \text{e} \qquad \int_{-\infty}^{\infty} f(x)\,dx = 1.$$

Um alerta que vale para sempre: $f(x)$ **não é** uma probabilidade — pode inclusive ser maior que 1. Probabilidade é área sob a densidade, não altura dela.

A **função de distribuição acumulada** no caso contínuo é

$$F(x_0) = P(X \le x_0) = \int_{-\infty}^{x_0} f(x)\,dx,$$

e dela seguem as identidades de trabalho:

$$P(a < X < b) = \int_a^b f(x)\,dx = F(b) - F(a), \qquad P(X < c) + P(X > c) = 1, \qquad P(X > c) = 1 - F(c).$$

3.2.1 A Normal e o papel da simetria

O exemplo canônico de distribuição contínua é a **Normal**, que reaparecerá em toda a inferência do curso — não por elegância, mas porque o teorema central do limite faz dela a distribuição-limite das médias amostrais e, por extensão, dos estimadores de MQO.

A Normal é **simétrica** em torno de sua média, e disso decorrem propriedades extras da CDF que o professor destaca. Para a Normal padrão,

$$P(X < -c) = P(X > c) \qquad \text{e portanto} \qquad F(-c) = 1 - F(c).$$

A primeira igualdade diz que as duas caudas, cortadas simetricamente, têm a mesma massa; a segunda a reescreve em termos da acumulada. A consequência prática é conhecida de quem já consultou uma tabela: **é por isso que os livros de estatística só trazem os valores positivos da CDF da Normal**. O lado negativo é redundante — obtém-se por $1 - F(c)$.

> **Ferramenta estatística — a distribuição Normal**
>
> As propriedades da Normal (forma da densidade, padronização, uso de tabelas, a família $\chi^2/t/F$ derivada dela) são tratadas em detalhe no capítulo de *Variáveis Aleatórias Contínuas* das **Notas de Aula de Estatística**. Aqui interessa apenas o que será usado adiante: simetria, padronização e o fato de que média e variância determinam completamente a distribuição.

3.3 Valor esperado

O **valor esperado** de uma variável aleatória é a média ponderada de seus valores possíveis, com pesos iguais às respectivas probabilidades. É o análogo populacional da média aritmética.

Para uma variável aleatória discreta,

$$\boxed{\mathbb{E}(X) = \sum_{i=1}^{n} p(x_i)\, x_i}$$

e para uma contínua, com a soma virando integral e a probabilidade virando densidade,

$$\boxed{\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x)\, dx}$$

Exemplo: $X \sim \mathbf{Bin}(n, \theta)$. Pode-se mostrar que o valor esperado de uma Binomial com parâmetros n e θ é

$$\mathbb{E}(X) = n\,\theta.$$

O resultado tem leitura imediata no exemplo dos pênaltis: um time com **80% de aproveitamento** vai converter, em média, $5 \times 0{,}8 = 4$ gols a cada 5 cobranças. Note que "em média" não significa

"sempre" — a Binomial atribui probabilidade positiva a todos os resultados de 0 a 5; significa que, repetida a disputa muitas vezes, a média de gols convertidos tende a 4. O laboratório verifica isso por simulação.

3.3.1 Propriedades

Duas propriedades do valor esperado carregam quase todo o trabalho algébrico do curso. A primeira é a **linearidade**: se $Y = a + bX$, então

$$\boxed{\mathbb{E}(Y) = a + b\,\mathbb{E}(X)}$$

A segunda trata de funções mais gerais: se $Y = g(X)$, então $\mathbb{E}(Y) = \mathbb{E}(g(X))$, que no caso contínuo se calcula como

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f(x)\,dx.$$

Repare no que essa fórmula diz — e no que ela não diz. Para calcular $\mathbb{E}(g(X))$ basta a densidade de X; não é preciso derivar a densidade de Y. Mas, em geral, $\mathbb{E}(g(X)) \neq g(\mathbb{E}(X))$: a esperança comuta com transformações **lineares** e apenas com elas. Essa distinção reaparecerá adiante, quando mostrarmos que $\mathbb{E}(Y \mid X) = \mathbb{E}(Y)$ não implica independência.

3.4 Variância e desvio padrão

O valor esperado localiza a distribuição; a **variância** mede sua dispersão em torno desse centro. Com $\mu = \mathbb{E}(X)$, no caso discreto,

$$\mathrm{Var}(X) = \sum_{i=1}^{n} p(x_i)\,(x_i - \mu)^2 \qquad \text{ou, equivalentemente,} \qquad \mathrm{Var}(X) = \sum_{i=1}^{n} p(x_i)\,x_i^2 \;-\; \mu^2,$$

e no caso contínuo,

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx \qquad \text{ou} \qquad \mathrm{Var}(X) = \int_{-\infty}^{\infty} x^2 f(x)\,dx \;-\; \mu^2.$$

As duas formas de cada par são idênticas — a segunda é o **atalho de cálculo** $\mathrm{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$, que se obtém expandindo o quadrado e usando a linearidade:

$$\mathbb{E}\big[(X-\mu)^2\big] = \mathbb{E}\big[X^2 - 2\mu X + \mu^2\big] = \mathbb{E}(X^2) - 2\mu\,\mathbb{E}(X) + \mu^2 = \mathbb{E}(X^2) - \mu^2.$$

Vale memorizar a forma compacta, que usaremos sem cerimônia:

$$\boxed{\mathrm{Var}(X) = \mathbb{E}\big[(X-\mu)^2\big] = \mathbb{E}(X^2) - \big[\mathbb{E}(X)\big]^2}$$

O **desvio padrão** é a raiz quadrada da variância:

$$\sigma = \mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}.$$

Por que trabalhar com o desvio padrão se a variância já mede dispersão? O professor dá dois motivos, e ambos importam.

Primeiro motivo: unidades. A variância está na unidade **ao quadrado** da variável. Se X é o retorno de um ativo financeiro medido em %, a unidade da variância é $\%^2$ — algo que ninguém sabe interpretar. O desvio padrão está em %, a unidade original, e por isso é a medida de risco que se reporta na prática.

Segundo motivo: a desigualdade de Chebyshev, que dá ao desvio padrão um significado quantitativo válido para *qualquer* distribuição.

3.4.1 A desigualdade de Chebyshev

O professor apresenta o resultado por suas implicações práticas: para uma ampla gama de distribuições de probabilidade, **75% dos valores** se encontram no intervalo de dois desvios padrão em torno da média, e **88,9%** dentro de três desvios padrão. Para comparação, no caso da distribuição Normal essas porcentagens são **95%** e **99,7%**.

Vale enunciar o resultado formalmente, porque a força dele está exatamente no que ele *não* exige.

! Desigualdade de Chebyshev

Seja X uma variável aleatória com média μ e variância $\sigma^2 < \infty$. Então, para todo $k > 0$,

$$\boxed{P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2} \qquad \Longleftrightarrow \qquad P(|X-\mu| < k\sigma) \geq 1 - \frac{1}{k^2}}$$

Nenhuma hipótese sobre a forma da distribuição é necessária — apenas que a variância exista.

Com $k = 2$, o limite é $1 - 1/4 = 0{,}75$; com $k = 3$, é $1 - 1/9 \approx 0{,}889$. São exatamente os 75% e os 88,9% do slide.

A demonstração é curta e instrutiva. Escreva a variância como integral (o argumento no caso discreto é idêntico, com somas) e jogue fora a parte central do domínio, que só pode contribuir com massa não negativa:

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx \;\geq\; \int_{|x-\mu|\geq k\sigma} (x-\mu)^2 f(x)\,dx.$$

Na região que sobrou vale $(x-\mu)^2 \geq k^2\sigma^2$ por construção. Substituindo o integrando por essa cota inferior, a desigualdade se preserva:

$$\sigma^2 \;\geq\; k^2\sigma^2 \int_{|x-\mu|\geq k\sigma} f(x)\,dx \;=\; k^2\sigma^2\, P(|X-\mu| \geq k\sigma).$$

Dividindo ambos os lados por $k^2\sigma^2 > 0$, obtemos $P(|X - \mu| \geq k\sigma) \leq 1/k^2$, como queríamos. ■

Duas leituras deste resultado, ambas importantes para o curso.

A primeira é **positiva**: o desvio padrão não é uma medida arbitrária de dispersão. Ele carrega, universalmente, uma garantia probabilística — dizer que σ é pequeno é dizer algo com conteúdo sobre onde a variável cai, sem precisar saber qual é a distribuição.

A segunda é sobre **folga**. Compare 75% (Chebyshev) com 95% (Normal): o limite de Chebyshev é bastante frouxo para distribuições bem comportadas. Isso é inevitável e é o preço da generalidade — o limite tem de valer inclusive para as distribuições mais patológicas com variância finita. Na prática, quando conhecemos a forma da distribuição, usamos o resultado mais forte que ela permite; Chebyshev é a rede de segurança de quando não conhecemos nada. O laboratório ilustra essa folga em cinco distribuições muito diferentes entre si.

3.4.2 Transformação linear e padronização

Seja $Y = a + bX$. Então

$$\boxed{\mathrm{Var}(Y) = b^2\,\mathrm{Var}(X) \qquad \text{e} \qquad \mathrm{SD}(Y) = |b|\,\mathrm{SD}(X)}$$

Duas coisas a notar. A constante aditiva a **desaparece**: deslocar toda a distribuição não altera sua dispersão. E o fator multiplicativo entra **ao quadrado** na variância, o que é apenas a contrapartida do problema das unidades — se b está em reais por dólar, a variância traz esse fator duas vezes. No desvio padrão aparece $|b|$, com módulo, porque o desvio padrão é não negativo por definição, mesmo quando $b < 0$.

Um caso particular merece nome próprio. Considere X com média μ_X e variância σ_X^2, e a transformação linear

$$Z = \frac{X - \mu_X}{\sigma_X},$$

que é a **padronização** de X. Aplicando as regras acima com $a = -\mu_X/\sigma_X$ e $b = 1/\sigma_X$, é fácil mostrar que

$$\mathbb{E}(Z) = 0 \qquad \text{e} \qquad \mathrm{Var}(Z) = 1.$$

A padronização é o que permite expressar qualquer variável em "número de desvios padrão de distância da média" e, com isso, comparar variáveis medidas em unidades diferentes. É também o mecanismo por trás de toda estatística de teste que veremos: uma estatística t é um estimador padronizado.

3.5 Distribuição conjunta, marginal e independência

Até aqui, uma variável de cada vez. Mas econometria é sobre **relações** entre variáveis, e para falar delas precisamos da distribuição de duas variáveis simultaneamente.

Caso discreto. A **distribuição conjunta** de X e Y é

$$P_{X,Y}(x,y) = P(X = x,\, Y = y),$$

a probabilidade de os dois eventos ocorrerem juntos. Para um conjunto S de pares,

$$P\big((X,Y) \in S\big) = \sum_{(x,y)\in S} P(x,y),$$

e as restrições são as análogas do caso univariado:

$$P_{X,Y}(x,y) \geq 0 \qquad \text{e} \qquad \sum_x \sum_y P(x,y) = 1.$$

A **distribuição marginal** de cada variável se obtém "somando fora" a outra:

$$P(X = x) = \sum_y P(x,y) \qquad \text{e} \qquad P(Y = y) = \sum_x P(x,y).$$

O nome vem literalmente da prática de anotar essas somas nas margens de uma tabela de dupla entrada — é o que faremos na tabela da próxima seção. A esperança de uma função de ambas as variáveis é

$$\mathbb{E}\big(g(X,Y)\big) = \sum_x \sum_y g(x,y)\, P(x,y).$$

Finalmente, X e Y são **independentes** se

$$\boxed{P(x,y) = P(X = x)\, P(Y = y), \qquad \forall\, x, y}$$

isto é, se a conjunta se fatora como o produto das marginais **para todos os pares**. A quantificação universal importa: basta uma célula em que a fatoração falhe para que a independência esteja destruída.

Caso contínuo. As definições são as mesmas, com densidades no lugar de probabilidades e integrais no lugar de somas:

$$P\big((X,Y) \in S\big) = \iint_S f(x,y)\, dx\, dy, \qquad f(x,y) \geq 0, \qquad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\, dx\, dy = 1.$$

As marginais são

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy \qquad \text{e} \qquad f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx,$$

a independência é $f(x,y) = f_X(x)\,f_Y(y)$ para todo (x,y), e a esperança de uma função $g(\cdot)$ é

$$\mathbb{E}(g(X,Y)) = \iint g(x,y)\,f(x,y)\,dx\,dy.$$

3.6 Distribuição condicional e média condicional

Chegamos ao coração do capítulo. Tudo o que veio antes é infraestrutura; o que vem a seguir é o objeto que a regressão estima.

A distribuição de uma variável aleatória Y **condicionada** a outra variável aleatória X, em um valor específico de X, é chamada **distribuição condicional de Y dado X**. A probabilidade condicional é

$$\boxed{P(Y = y \mid X = x) = \frac{P(X = x,\, Y = y)}{P(X = x)}}$$

A leitura é a de sempre: restringimos o universo aos casos em que $X = x$ — daí o denominador — e perguntamos qual fração deles tem também $Y = y$. Condicionar é **reponderar**: descartamos parte do espaço amostral e renormalizamos o que sobra para que volte a somar 1.

3.6.1 A tabela chuva × tempo de deslocamento

O exemplo do professor é o de um trabalhador que se desloca de casa para o trabalho. Sejam $X = 0$ se chove e $X = 1$ se não chove, e $Y = 0$ se o deslocamento é longo (*long commute*) e $Y = 1$ se é curto (*short commute*). A distribuição conjunta é:

Tabela 1: Distribuição conjunta de chuva e tempo de deslocamento. As quatro células internas são a conjunta; as margens são as distribuições marginais.

	Rain ($X = 0$)	No Rain ($X = 1$)	**Total**
Long Commute ($Y = 0$)	0,15	0,07	**0,22**
Short Commute ($Y = 1$)	0,15	0,63	**0,78**
Total	**0,30**	**0,70**	**1,00**

As quatro células internas somam 1, como exige a definição de distribuição conjunta. As margens são as marginais: somando a linha, $P(Y = 0) = 0{,}15 + 0{,}07 = 0{,}22$; somando a coluna, $P(X = 0) = 0{,}15 + 0{,}15 = 0{,}30$. Chove em 30% dos dias, e o deslocamento é longo em 22% deles.

Os cálculos do professor, um a um:

$$P(\text{Long C. \& Rain}) = P(X = 0, Y = 0) = 0{,}15,$$

que é a **conjunta** — lida diretamente da célula;

$$P(\text{Long C.}) = P(Y = 0) = P(X = 0, Y = 0) + P(X = 1, Y = 0) = 0{,}15 + 0{,}07 = 0{,}22,$$

que é a **marginal** — a soma da linha; e as **condicionais**, que dividem a conjunta pela marginal do condicionante:

$$P(\text{Long C.} \mid \text{Rain}) = \frac{P(\text{Long C. \& Rain})}{P(\text{Rain})} = \frac{P(X = 0, Y = 0)}{P(X = 0)} = \frac{0{,}15}{0{,}30} = 0{,}5,$$

$$P(\text{Short C.} \mid \text{Rain}) = \frac{P(X = 0, Y = 1)}{P(X = 0)} = \frac{0{,}15}{0{,}30} = 0{,}5.$$

Vale contrastar os três números que aparecem para o mesmo evento "deslocamento longo": 0,15 (conjunta com chuva), 0,22 (marginal) e 0,50 (condicional a chuva). São perguntas diferentes. A conjunta pergunta "que fração de *todos* os dias tem chuva **e** deslocamento longo?"; a marginal, "que fração de todos os dias tem deslocamento longo?"; a condicional, "**entre os dias de chuva**, que fração tem deslocamento longo?". Confundir a segunda com a terceira é um dos erros mais comuns de leitura de dados — e, note, o salto de 22% para 50% é precisamente a informação econômica que a tabela contém: chover mais que dobra a chance de um deslocamento longo.

Observe também que a condicional em dia de chuva é (0,5; 0,5), enquanto em dia sem chuva é (0,07/0,70; 0,63/0,70) = (0,1; 0,9). Como as duas condicionais diferem, X e Y **não são independentes** — o que também se verifica pela fatoração: $P(X = 0)P(Y = 0) = 0{,}30 \times 0{,}22 = 0{,}066 \neq 0{,}15$.

3.6.2 A média condicional

A **média condicional** (*conditional expectation*) de Y dado X é a média da distribuição condicional de Y dado X. Ou seja: é o valor esperado de Y calculado usando a distribuição condicional, e não a marginal. Se Y assume k valores $y_1, y_2, \ldots, y_k$,

$$\boxed{\mathbb{E}(Y \mid X = x) = \sum_{i=1}^{k} y_i \, P(Y = y_i \mid X = x)}$$

Duas observações de notação que evitam confusão mais adiante. Primeiro, $\mathbb{E}(Y \mid X = x)$ é um **número** — a média de Y na subpopulação em que X vale x. Segundo, $\mathbb{E}(Y \mid X)$, sem fixar o valor, é uma **função de** X e, portanto, uma variável aleatória: ela assume o valor $\mathbb{E}(Y \mid X = x)$ quando $X = x$. Essa distinção é o que torna sensata a expressão $\mathbb{E}(\mathbb{E}(Y \mid X))$ da próxima seção — a esperança externa é tomada sobre a aleatoriedade de X.

Na tabela do deslocamento, com Y codificada como 0/1, $\mathbb{E}(Y \mid X = 0) = 0 \times 0{,}5 + 1 \times 0{,}5 = 0{,}5$ e $\mathbb{E}(Y \mid X = 1) = 0 \times 0{,}1 + 1 \times 0{,}9 = 0{,}9$. Para variáveis binárias, a média condicional é simplesmente a probabilidade condicional de $Y = 1$ — fato que reaparecerá no modelo de probabilidade linear.

! Por que a média condicional é o objeto central do curso

O modelo de regressão populacional de 1 será $Y_i = \beta_0 + \beta_1 X_i + u_i$ com $\mathbb{E}(u_i \mid X_i) = 0$. Tomando a média condicional dos dois lados e usando as propriedades abaixo,

$$\mathbb{E}(Y_i \mid X_i) = \beta_0 + \beta_1 X_i.$$

Isto é: a reta populacional **é** a média condicional de Y dado X, e β_1 é o quanto essa média muda quando X aumenta em uma unidade. Estimar uma regressão é estimar médias condicionais.

3.6.3 Propriedades da média condicional

Três propriedades, todas usadas repetidamente adiante:

$$\mathbb{E}(g(X) \mid X) = g(X);$$

$$\mathbb{E}(a(X)\,Y + b(X) \mid X) = a(X)\,\mathbb{E}(Y \mid X) + b(X);$$

$$\text{se } X \text{ e } Y \text{ são independentes}, \qquad \mathbb{E}(Y \mid X) = \mathbb{E}(Y).$$

A primeira formaliza a ideia de que, **dado** X, qualquer função de X é conhecida — e o valor esperado de uma constante é a própria constante. A segunda é a linearidade condicional, e a novidade em relação à linearidade usual é que os "coeficientes" $a(X)$ e $b(X)$ podem depender de X: condicionalmente a X, eles são constantes e saem da esperança. A terceira diz que, sob independência, saber X não informa nada sobre Y — a distribuição condicional coincide com a marginal, e portanto suas médias coincidem.

Guarde a terceira propriedade com atenção à direção da implicação: independência **implica** $\mathbb{E}(Y \mid X) = \mathbb{E}(Y)$. A recíproca é falsa, e a seção final deste capítulo é inteiramente dedicada a isso.

3.7 A lei das expectativas iteradas

Enunciada em palavras, a **lei das expectativas iteradas** diz que a média de Y é a média ponderada da média condicional de Y dado X, com pesos iguais à distribuição de probabilidade de X. Matematicamente:

$$\boxed{\mathbb{E}(Y) = \sum_{i=1}^{n} \mathbb{E}(Y \mid X = x_i) \cdot \Pr(X = x_i) \;=\; \mathbb{E}(\mathbb{E}(Y \mid X))}$$

A segunda igualdade é apenas a primeira reescrita: a soma ponderada de $\mathbb{E}(Y \mid X = x_i)$ pelas probabilidades de X é, por definição, a esperança da variável aleatória $\mathbb{E}(Y \mid X)$.

3.7.1 A intuição, pela altura média

O professor constrói a intuição com um problema concreto, e a passagem é boa o bastante para reproduzir passo a passo.

O problema: calcular a altura média de uma população.

Primeira forma. Simplesmente calcular a média de todo mundo:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i,$$

que é o equivalente amostral de $\mathbb{E}(Y)$. Nada de especial aqui.

Segunda forma. Separar a população entre mulheres (M) e homens (H) e calcular a altura média de cada grupo:

$$\bar{y}_M = \frac{1}{n_M} \sum_{i \in M} y_i \qquad \text{e} \qquad \bar{y}_H = \frac{1}{n_H} \sum_{i \in H} y_i,$$

com $n_M + n_H = n$. A pergunta é: podemos recuperar a média geral a partir dessas duas médias parciais? Podemos, e o caminho é curto.

Multiplique cada média de grupo pelo tamanho do grupo. Isso desfaz a divisão e devolve as **somas** de cada grupo, que juntas são a soma total:

$$n_M \, \bar{y}_M + n_H \, \bar{y}_H = \sum_{i \in M} y_i + \sum_{i \in H} y_i = \sum_{i=1}^{n} y_i.$$

Basta então dividir ambos os lados por n:

$$\bar{y} = \frac{n_M}{n} \, \bar{y}_M + \frac{n_H}{n} \, \bar{y}_H.$$

Agora o passo que transforma uma identidade contábil em um teorema de probabilidade. Observe que

$$\Pr(M) = \frac{n_M}{n} \qquad \text{e} \qquad \Pr(H) = \frac{n_H}{n},$$

isto é, as frações populacionais **são** as probabilidades de sexo. E as médias de grupo são médias condicionais: com X denotando o sexo,

$$\mathbb{E}(Y \mid X = m) = \bar{y}_M \qquad \text{e} \qquad \mathbb{E}(Y \mid X = h) = \bar{y}_H.$$

Substituindo tudo na identidade anterior, com notação de esperanças:

$$\mathbb{E}(Y) = \mathbb{E}(Y \mid X = m) \cdot \Pr(X = m) \; + \; \mathbb{E}(Y \mid X = h) \cdot \Pr(X = h),$$

ou, compactamente,

$$\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y \mid X)).$$

Essa é a lei das expectativas iteradas. Note o que ela realmente afirma: a média incondicional não é uma quantidade nova, independente das médias condicionais — ela é a **média ponderada** delas. Nada se perde ao estratificar a população, desde que os pesos usados na recomposição sejam as probabilidades corretas de cada estrato. E note também o que ela **não** afirma: a média simples das médias condicionais, $(\bar{y}_M + \bar{y}_H)/2$, em geral **não** é $\bar{y}$ — só coincide se os grupos tiverem o mesmo tamanho. Ignorar os pesos é uma fonte clássica de erro em análise de dados agregados.

Duas consequências que usaremos muito. A primeira: se $\mathbb{E}(Y \mid X) = 0$ para todo X, então $\mathbb{E}(Y) = 0$ — a lei transporta a hipótese de média condicional nula para a média incondicional. É exatamente o mecanismo pelo qual a hipótese $\mathbb{E}(u_i \mid X_i) = 0$ garante que os resíduos populacionais têm média zero. A segunda, mais geral, é a **lei das expectativas iteradas em forma multiplicativa**: para qualquer função g,

$$\mathbb{E}(g(X)\,Y) = \mathbb{E}(g(X)\,\mathbb{E}(Y \mid X)),$$

que se obtém aplicando a lei a $g(X)Y$ e usando a linearidade condicional. Com $g(X) = X$ e $\mathbb{E}(Y \mid X)$ constante, ela produzirá, na próxima seção, o resultado sobre covariância nula.

3.8 Covariância, correlação e independência

A **covariância** é uma medida de associação **linear** entre duas variáveis aleatórias:

$$\boxed{\operatorname{Cov}(X, Y) = \sigma_{XY} = \mathbb{E}\big[(X - \mu_X)(Y - \mu_Y)\big]}$$

A lógica do sinal é a de sempre: cada parcela é positiva quando as duas variáveis se desviam de suas médias no mesmo sentido, negativa quando em sentidos opostos. Se predominam os pares concordantes, $\operatorname{Cov} > 0$.

Se X e Y são **independentes**, então $\operatorname{Cov}(X, Y) = 0$. A demonstração é imediata: sob independência, $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$, e expandindo o produto na definição obtém-se $\operatorname{Cov}(X, Y) = \mathbb{E}(XY) - \mu_X\mu_Y = 0$.

E a recíproca? O professor formula a pergunta e a deixa no ar: se $\operatorname{Cov}(X, Y) = 0$, então X e Y são independentes? **Não.** Por quê?

Porque a covariância mede apenas **dependência linear**. Ela é construída somando produtos de desvios, e uma dependência que seja simétrica em torno da média produz parcelas positivas e negativas que se cancelam exatamente — mesmo que a dependência seja total.

Contraexemplo concreto. Seja X uma variável aleatória com distribuição simétrica em torno de zero — por exemplo, $X \sim \text{Uniforme}(-2, 2)$ — e defina

$$Y = X^2.$$

Não há dependência mais completa: conhecendo X, conhecemos Y com certeza absoluta. E, no entanto,

$$\mathrm{Cov}(X, Y) = \mathbb{E}\big[(X - \mu_X)(Y - \mu_Y)\big] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}(X^3) - 0 = 0,$$

porque $\mathbb{E}(X) = 0$ e, pela simetria da distribuição de X, todos os momentos ímpares são nulos: $\mathbb{E}(X^3) = 0$. A covariância — e portanto a correlação — é **exatamente zero** entre duas variáveis em que uma é função determinística da outra.

A moral: ausência de correlação é uma afirmação sobre a **inexistência de associação linear**, não sobre a inexistência de associação. O laboratório desenha esse contraexemplo, e o gráfico é a figura mais eloquente do capítulo — uma parábola perfeita com correlação amostral próxima de zero.

3.8.1 O problema das unidades e a correlação

O **sinal** da covariância é informativo: sabemos se as variáveis andam juntas na mesma direção ($\mathrm{Cov} > 0$) ou em direções opostas ($\mathrm{Cov} < 0$). Mas sua **magnitude** é problemática, por duas razões relacionadas.

A primeira é a unidade de medida. Se calculamos a covariância entre peso e altura, sua unidade é quilograma-metro — algo difícil de interpretar e impossível de comparar com a covariância entre, digamos, renda e escolaridade.

A segunda, mais séria, é que a covariância **não é robusta a transformações lineares**. Sejam $W = aX + b$ e $Z = cY + d$. Então

$$\boxed{\mathrm{Cov}(W, Z) = \mathrm{Cov}(aX + b,\ cY + d) = ac \cdot \mathrm{Cov}(X, Y)}$$

As constantes aditivas somem — deslocar as variáveis não altera sua co-variação —, mas os fatores multiplicativos passam direto. Medir a renda em centavos em vez de reais multiplica a covariância por cem, sem que nada tenha mudado na relação entre as variáveis. A covariância é sensível à unidade de medida, e isso a inutiliza como medida de **força** de associação.

A solução é padronizar. A **correlação** é a covariância dividida pelo produto dos desvios padrão:

$$\boxed{\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}}$$

Como os desvios padrão carregam as mesmas unidades que a covariância — uma vez cada —, o quociente é **adimensional**. E como se pode mostrar (é a desigualdade de Cauchy–Schwarz), ele é limitado:

$$-1 \le \mathrm{Corr}(X, Y) \le 1.$$

Os extremos correspondem à relação linear perfeita, e $\rho = 0$ à ausência de associação *linear*. Verifique a robustez: aplicando a regra acima ao numerador e a regra $\mathrm{SD}(aX+b) = |a|\mathrm{SD}(X)$ ao denominador, o fator $|ac|$ se cancela e a correlação fica inalterada (a menos do sinal, se $ac < 0$).

3.8.2 Variância de combinações lineares

Duas propriedades que serão usadas o tempo todo — na dedução da variância do estimador de MQO, entre outros lugares:

$$\boxed{\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X, Y)}$$

$$\boxed{\mathrm{Var}(aX - bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) - 2ab\,\mathrm{Cov}(X, Y)}$$

A segunda é a primeira com b trocado por $-b$. O termo de covariância é o que distingue o caso geral do caso independente: sob independência ele desaparece e as variâncias simplesmente se somam. Em finanças, essa é a álgebra inteira da diversificação — combinar ativos com covariância negativa reduz a variância da carteira abaixo da média das variâncias individuais.

3.9 Correlação e média condicional: uma distinção que importa

O professor dedica três slides a um ponto sutil, e não é por acaso: é exatamente essa distinção que fundamenta a hipótese central do próximo capítulo. Vale destrinchá-la em três afirmações separadas.

Primeira: se a média condicional de Y não depende de X, então X e Y são não correlacionados. Ou seja,

$$\mathbb{E}(Y \mid X) = \mathbb{E}(Y) \implies \mathrm{Cov}(X, Y) = 0 \;\text{ e }\; \mathrm{Corr}(X, Y) = 0.$$

A demonstração usa a lei das expectativas iteradas em forma multiplicativa. Escreva $\mathbb{E}(XY) = \mathbb{E}(X\,\mathbb{E}(Y \mid X))$; se $\mathbb{E}(Y \mid X) = \mathbb{E}(Y) = \mu_Y$ é constante, ela sai da esperança:

$$\mathbb{E}(XY) = \mu_Y\,\mathbb{E}(X) = \mu_X\mu_Y \quad \implies \quad \mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mu_X\mu_Y = 0.$$

Segunda: a recíproca não vale. Não é necessariamente verdade que, se X e Y são não correlacionados, então a média condicional de Y dado X não depende de X. Ou, alternativamente: é perfeitamente possível que $\mathbb{E}(Y \mid X)$ **dependa** de X e ainda assim a correlação entre Y e X seja zero.

O contraexemplo já está construído: com X simétrica em torno de zero e $Y = X^2$, temos $\mathrm{Corr}(X, Y) = 0$ mas $\mathbb{E}(Y \mid X) = X^2$, que obviamente depende de X — e depende de forma dramática. A relação existe, é determinística, e a correlação não a enxerga porque não é linear.

Terceira: mesmo $\mathbb{E}(Y \mid X) = \mathbb{E}(Y)$ não implica independência. Este é o ponto mais fino dos três, e o professor o registra explicitamente: não é verdade que Y e X são independentes quando a média condicional de Y dado X não depende de X. Isso indica apenas que Y e X não estão correlacionados linearmente, ou que **não há dependência na média**. Ainda pode haver dependência em termos de **variância**, de **distribuição** ou de outras relações não lineares.

Um exemplo torna isso palpável. Suponha $X \sim \mathrm{Uniforme}(-1, 1)$ e $Y = X \cdot \varepsilon$, com ε independente de X, $\mathbb{E}(\varepsilon) = 0$ e $\mathrm{Var}(\varepsilon) = 1$. Então

$$\mathbb{E}(Y \mid X) = X\,\mathbb{E}(\varepsilon) = 0 = \mathbb{E}(Y)$$

— a média condicional é constante, não depende de X. Mas

$$\mathrm{Var}(Y \mid X) = X^2\,\mathrm{Var}(\varepsilon) = X^2,$$

que depende de X. Conhecer X não muda a previsão pontual de Y, mas muda completamente a **incerteza** dessa previsão. As variáveis são dependentes, e a dependência é toda na variância. (Este exemplo é, aliás, a definição de **heterocedasticidade**, que reaparecerá quando discutirmos erros-padrão robustos.)

Vale organizar as três afirmações em uma hierarquia, da hipótese mais forte para a mais fraca:

$$\underbrace{X \perp Y}_{\text{independência}} \implies \underbrace{\mathbb{E}(Y \mid X) = \mathbb{E}(Y)}_{\text{independência em média}} \implies \underbrace{\mathrm{Corr}(X, Y) = 0}_{\text{ausência de correlação}}$$

com **nenhuma** das setas valendo no sentido inverso.

> **!** Por que isto importa em Seção 4
>
> A hipótese fundamental do modelo de regressão é $\mathbb{E}(u_i \mid X_i) = 0$ — a do **meio** da hierarquia acima, não a da direita. Ela é **estritamente mais forte** que $\mathrm{Corr}(u_i, X_i) = 0$: exige que o erro não tenha *nenhuma* relação em média com o regressor, em nenhum valor de X, e não apenas que a relação linear média seja nula.
> A diferença não é acadêmica. Se o verdadeiro modelo é quadrático e ajustamos uma reta, é possível que o resíduo seja não correlacionado com X por construção do MQO — a álgebra do estimador garante isso na amostra — e ainda assim $\mathbb{E}(u \mid X) \neq 0$ na população, com o efeito estimado sistematicamente errado em algumas faixas de X. Ausência de correlação é o que o MQO impõe mecanicamente; média condicional nula é o que precisa ser **verdade no mundo** para que a estimativa seja interpretável causalmente. E, como veremos, essa segunda condição **não é testável**.

3.10 Correlação e causalidade

O curso fecha esta aula onde abriu a anterior. O que é a correlação entre duas variáveis? É uma medida da força, da direção e do grau de **associação** entre elas: indica se as duas se movem juntas ao longo do tempo ou de uma amostra, de modo que o aumento ou a diminuição de uma esteja associado a uma mudança na outra.

Tudo o que este capítulo construiu deveria deixar claro o que a correlação **não** é. Ela não distingue quem causa quem — $\mathrm{Corr}(X, Y) = \mathrm{Corr}(Y, X)$, a medida é perfeitamente simétrica, enquanto causalidade não é. Ela não exclui a possibilidade de uma terceira variável estar causando ambas. E, como vimos na seção anterior, ela sequer capta toda a associação existente: pode ser zero entre variáveis deterministicamente ligadas.

> ❗ A frase do curso
>
> **Correlação NÃO implica causalidade.**
> O Seção 2 mostrou o custo de esquecê-la — em terapia de reposição hormonal, em Vioxx, em talidomida. Este capítulo forneceu o vocabulário para dizer com precisão o que falta: uma correlação é uma afirmação sobre a distribuição conjunta observada, e um efeito causal é uma afirmação sobre o que aconteceria sob **intervenção**. Os capítulos seguintes são, do começo ao fim, uma sequência de estratégias para vencer essa distância.

3.11 Laboratório computacional em R

Quatro exercícios computacionais, um por bloco do capítulo. Todos rodam com R base e `ggplot2`/`dplyr`/`patchwork`, e os números impressos são os que o R de fato produziu na compilação deste PDF.

3.11.1 A Binomial dos pênaltis

Começamos verificando, célula a célula, a anatomia da fórmula binomial: quantas sequências produzem cada total, qual a probabilidade de *uma* sequência, e o produto das duas.

```
theta <- 0.8; n <- 5
x <- 0:n
dist_pen <- data.frame(
  x           = x,
  combinacoes = choose(n, x),                       # C(n, x): quantas sequencias
  sequencia   = theta^x * (1 - theta)^(n - x),      # prob. de UMA sequencia
  prob        = choose(n, x) * theta^x * (1 - theta)^(n - x),
  dbinom      = dbinom(x, n, theta),                # a funcao nativa, para conferir
  acumulada   = pbinom(x, n, theta)
)
round(dist_pen, 5)
```

```
  x combinacoes sequencia    prob  dbinom acumulada
1 0           1   0.00032 0.00032 0.00032   0.00032
2 1           5   0.00128 0.00640 0.00640   0.00672
3 2          10   0.00512 0.05120 0.05120   0.05792
4 3          10   0.02048 0.20480 0.20480   0.26272
5 4           5   0.08192 0.40960 0.40960   0.67232
6 5           1   0.32768 0.32768 0.32768   1.00000
```

A coluna `sequencia` é a armadilha do slide e a coluna `prob` é a resposta correta. Para $x = 3$:

```
c(uma_sequencia = theta^3 * (1 - theta)^2,
  n_sequencias  = choose(5, 3),
  P_X_igual_3   = choose(5, 3) * theta^3 * (1 - theta)^2)
```

```
uma_sequencia  n_sequencias   P_X_igual_3
      0.02048      10.00000       0.20480
```

A probabilidade de uma sequência específica como $(1, 1, 1, 0, 0)$ é 0,02048; como existem $C_3^5 = 10$ sequências distintas com três gols, $P(X = 3) = 0{,}2048$ — **dez vezes maior**. Ignorar o coeficiente binomial erra por uma ordem de grandeza.

Agora $\mathbb{E}(X) = n\theta$, verificado por simulação: sorteamos 200 mil disputas de cinco cobranças, cada cobrança uma Bernoulli(0,8) independente, e contamos os gols.

```
set.seed(20260810)
sorteios <- matrix(rbinom(200000 * n, size = 1, prob = theta), ncol = n)
convertidos <- rowSums(sorteios)          # gols em cada disputa de 5 cobrancas
round(c(media_simulada = mean(convertidos),   n_theta        = n * theta,
        var_simulada   = var(convertidos),    n_theta_1theta = n * theta * (1 - theta)),
      4)
```

```
media_simulada        n_theta   var_simulada n_theta_1theta
        3.9974         4.0000         0.8030         0.8000
```

A média simulada bate com $n\theta = 4$: o time com 80% de aproveitamento converte, em média, 4 dos 5 pênaltis. A variância simulada bate com $n\theta(1-\theta) = 0{,}8$.

```
library(patchwork)
g_pmf <- ggplot(dist_pen, aes(factor(x), prob)) +
  geom_col(fill = az, alpha = 0.75, width = 0.65) +
  geom_text(aes(label = sprintf("%.3f", prob)), vjust = -0.4,
            size = 3, colour = cz) +
  scale_y_continuous(expand = expansion(mult = c(0, 0.15))) +
  labs(x = "gols convertidos em 5 cobrancas", y = "P(X = x)")
g_cdf <- ggplot(dist_pen, aes(x, acumulada)) +
  geom_step(colour = vd, linewidth = 0.9, direction = "hv") +
  geom_point(colour = vd, size = 2) +
  scale_x_continuous(breaks = 0:5) +
  labs(x = "x", y = "F(x) = P(X <= x)")
g_pmf + g_cdf
```

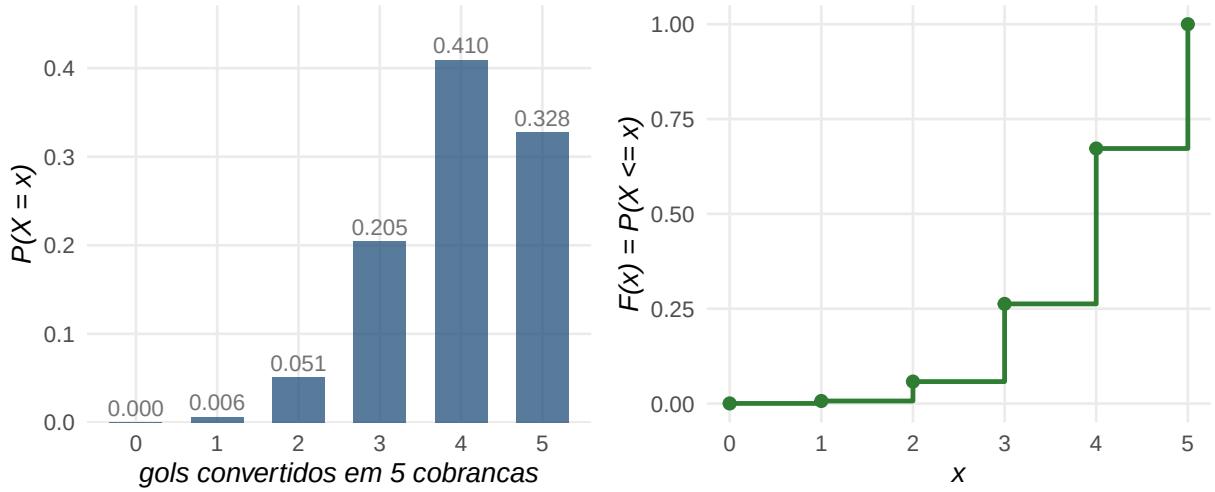

Figura 2: Distribuição Binomial(5; 0,8) dos pênaltis convertidos. À esquerda, a função de probabilidade; à direita, a acumulada — que nunca decresce e atinge 1 no maior valor possível.

3.11.2 Chebyshev em cinco distribuições

A desigualdade de Chebyshev vale para *qualquer* distribuição com variância finita. Para apreciar tanto sua universalidade quanto sua folga, aplicamos o teste a cinco distribuições deliberadamente diferentes entre si: uma simétrica de caudas leves (Normal), uma assimétrica (Exponencial), uma de suporte limitado (Uniforme), uma de caudas pesadas (t com 4 graus de liberdade) e uma assimétrica de suporte positivo (χ^2_2).

```
set.seed(1)
N <- 500000
amostras <- list(
  Normal      = rnorm(N),
  Exponencial = rexp(N, rate = 1),
  Uniforme    = runif(N, -1, 1),
  t_4gl       = rt(N, df = 4),
  Qui2_2gl    = rchisq(N, df = 2)
)
dentro <- function(v, k) mean(abs(v - mean(v)) <= k * sd(v))
cheby <- data.frame(
  distribuicao = names(amostras),
  k2_simulado  = round(sapply(amostras, dentro, k = 2), 4),
  k2_limite    = 1 - 1/4,
  k3_simulado  = round(sapply(amostras, dentro, k = 3), 4),
  k3_limite    = round(1 - 1/9, 4),
  row.names = NULL
)
cheby
```

```
  distribuicao k2_simulado k2_limite k3_simulado k3_limite
1       Normal      0.9544      0.75      0.9974    0.8889
2  Exponencial      0.9499      0.75      0.9818    0.8889
3     Uniforme      1.0000      0.75      1.0000    0.8889
4        t_4gl      0.9532      0.75      0.9871    0.8889
5     Qui2_2gl      0.9500      0.75      0.9819    0.8889
```

Duas leituras. Primeiro, **o limite nunca é violado**: em todas as cinco distribuições a fração simulada dentro de k desvios padrão excede $1 - 1/k^2$. Segundo, ele é **folgado** — a coluna simulada fica em torno de 95% para $k = 2$, contra a garantia de apenas 75%. A Uniforme é o caso extremo: seu suporte inteiro cabe dentro de dois desvios padrão, de modo que a fração é exatamente 1 contra um limite de 0,75.

```
c(violacoes_k2 = sum(cheby$k2_simulado < cheby$k2_limite),
  violacoes_k3 = sum(cheby$k3_simulado < cheby$k3_limite))
```

```
violacoes_k2 violacoes_k3
           0            0
```

Zero violações, como o teorema garante. A folga é o preço de uma garantia que não pede nada em troca: como o limite precisa valer inclusive para as distribuições mais adversas com variância finita, ele é necessariamente conservador para as bem comportadas.

3.11.3 A tabela chuva × commute e a lei das expectativas iteradas

Montamos a distribuição conjunta do professor e extraímos dela marginais, condicionais e a verificação numérica da lei das expectativas iteradas.

```
conj <- matrix(c(0.15, 0.07,
                 0.15, 0.63),
               nrow = 2, byrow = TRUE,
               dimnames = list(Y = c("Long (Y=0)", "Short (Y=1)"),
                               X = c("Rain (X=0)", "No Rain (X=1)")))
addmargins(conj)      # a tabela do slide, com as margens
```

```
             X
Y             Rain (X=0) No Rain (X=1)  Sum
  Long (Y=0)        0.15          0.07 0.22
  Short (Y=1)       0.15          0.63 0.78
  Sum               0.30          0.70 1.00
```

As margens reproduzem a tabela: 0,22 e 0,78 nas linhas, 0,30 e 0,70 nas colunas, 1,00 no canto. As condicionais de Y dado X saem dividindo cada coluna por sua marginal:

```r
marg_X <- colSums(conj)
marg_Y <- rowSums(conj)
cond_Y_dado_X <- sweep(conj, 2, marg_X, "/")   # divide cada coluna por P(X = x)
round(cond_Y_dado_X, 4)
```

```
             X
Y             Rain (X=0) No Rain (X=1)
  Long (Y=0)         0.5           0.1
  Short (Y=1)        0.5           0.9
```

A primeira coluna é (0,5; 0,5) — os dois cálculos do professor, $P(\text{Long C.} \mid \text{Rain})$ e $P(\text{Short C.} \mid \text{Rain})$. A segunda é (0,1; 0,9). Como as duas colunas diferem, X e Y não são independentes; note que $P(\text{Long C.}) = 0{,}22$ mas $P(\text{Long C.} \mid \text{Rain}) = 0{,}50$.

Codificando $Y \in \{0, 1\}$, as médias condicionais e a lei das expectativas iteradas:

```r
y <- c(0, 1)
E_Y_dado_X <- colSums(cond_Y_dado_X * y)   # E(Y | X = x) para cada x
E_Y_dado_X
```

```
   Rain (X=0) No Rain (X=1)
          0.5           0.9
```

```r
c(E_Y_direto     = sum(marg_Y * y),                 # E(Y) pela marginal
  E_de_E_Y_dado_X = sum(E_Y_dado_X * marg_X))       # E(E(Y|X)): media ponderada
```

```
     E_Y_direto E_de_E_Y_dado_X
           0.78            0.78
```

Os dois números coincidem em 0,78: a média incondicional de Y é a média ponderada das médias condicionais, com pesos 0,30 e 0,70 — as probabilidades de X. É a lei das expectativas iteradas em duas linhas de código, e é exatamente a álgebra da altura média por sexo, com "chove/não chove" no lugar de "mulher/homem".

Confirmando a dependência pela fatoração, e medindo a associação linear:

```r
produto_marginais <- outer(marg_Y, marg_X)          # o que a conjunta seria sob independencia
EX  <- sum(marg_X * c(0, 1)); EY <- sum(marg_Y * y)
EXY <- sum(outer(y, c(0, 1)) * conj)
round(c(conjunta_00 = conj[1, 1], sob_independencia = produto_marginais[1, 1],
        cov_XY = EXY - EX * EY), 4)
```

```
      conjunta_00 sob_independencia            cov_XY
            0.150             0.066             0.084
```

A célula (Long, Rain) vale 0,15, contra 0,066 que valeria sob independência. A covariância é positiva (0,084): não chover e ter deslocamento curto andam juntos.

3.11.4 O contraexemplo: covariância zero sem independência

Este é o resultado mais importante do laboratório. Geramos $X \sim \text{Uniforme}(-2, 2)$ — simétrica em torno de zero — e $Y = X^2$ mais um pequeno ruído, de modo que Y é, para todos os efeitos, uma função determinística de X.

```
set.seed(20260810)
M <- 20000
X <- runif(M, -2, 2)
Y <- X^2 + rnorm(M, sd = 0.25)
round(c(cov_XY = cov(X, Y), corr_XY = cor(X, Y)), 4)
```

```
 cov_XY corr_XY
-0.0103 -0.0073
```

A covariância e a correlação são **praticamente zero** — e o valor populacional é exatamente zero, porque $\text{Cov}(X, X^2) = \mathbb{E}(X^3) - \mathbb{E}(X)\mathbb{E}(X^2) = 0$ pela simetria de X. No entanto, a média condicional depende fortemente de X:

```
faixas <- cut(X, breaks = seq(-2, 2, by = 0.5))
round(tapply(Y, faixas, mean), 3)
```

```
(-2,-1.5] (-1.5,-1] (-1,-0.5]   (-0.5,0]    (0,0.5]    (0.5,1]    (1,1.5]    (1.5,2]
    3.077     1.568     0.585     0.080     0.091     0.591     1.584     3.091
```

$\mathbb{E}(Y \mid X)$ varia de cerca de 3,1 nas pontas a praticamente zero no centro. A relação é forte, sistemática e perfeitamente visível — e a correlação, cega para ela.

```
dd <- data.frame(X = X, Y = Y)
medias_cond <- dd |>
  mutate(faixa = cut(X, breaks = seq(-2, 2, by = 0.25))) |>
  group_by(faixa) |>
  summarise(x = mean(X), m = mean(Y), .groups = "drop")

set.seed(7)
dn <- data.frame(X = runif(M, -2, 2), Y = rnorm(M, mean = mean(Y), sd = sd(Y)))
medias_cond_n <- dn |>
  mutate(faixa = cut(X, breaks = seq(-2, 2, by = 0.25))) |>
  group_by(faixa) |>
  summarise(x = mean(X), m = mean(Y), .groups = "drop")

painel <- function(dados, medias, rotulo) {
  ggplot(dados, aes(X, Y)) +
    geom_point(alpha = 0.10, colour = cz, size = 0.6) +
    geom_hline(yintercept = mean(dados$Y), colour = az, linetype = "dashed") +
    geom_line(data = medias, aes(x, m), colour = vm, linewidth = 1) +
```

```
    labs(x = "X ~ Uniforme(-2, 2)", y = rotulo)
}
painel(dd, medias_cond, "Y = X^2 + ruido  (corr ~ 0)") +
  painel(dn, medias_cond_n, "Y independente de X  (corr ~ 0)")
```

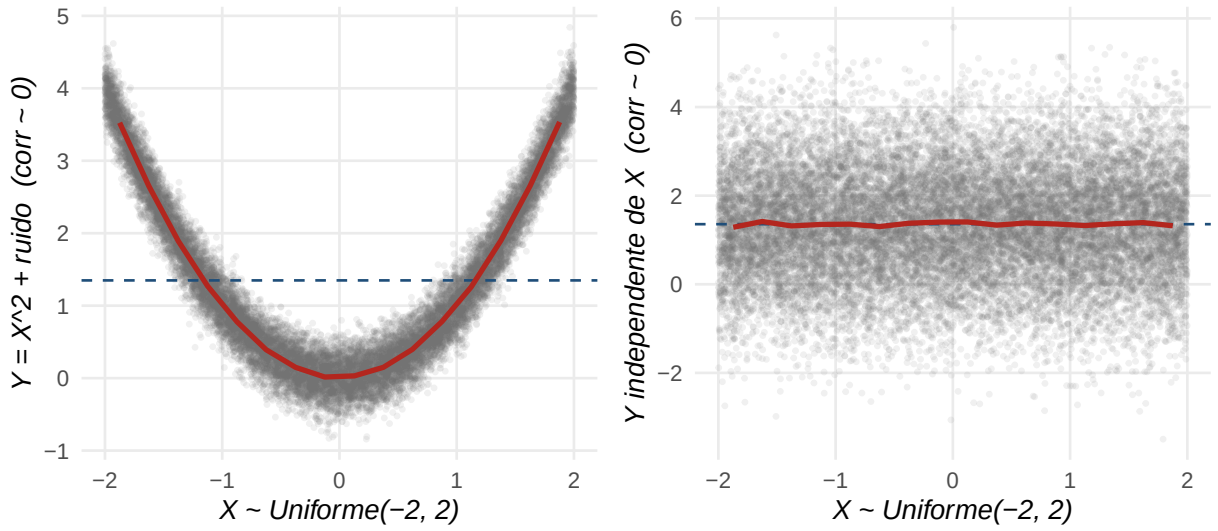

Figura 3: Correlação zero, dependência total. À esquerda, $Y = X^2 + \text{ruído}$ com X simétrica: a correlação amostral é ~0, mas a média condicional (linha vermelha) desenha uma parábola. À direita, Y genuinamente independente de X, para comparação: aí sim a média condicional é plana. A linha azul tracejada é a média incondicional $\mathbb{E}(Y)$ nos dois painéis.

Os dois painéis têm correlação aproximadamente zero. No da direita, Y é genuinamente independente de X e a média condicional coincide com a incondicional — as linhas vermelha e azul se sobrepõem. No da esquerda, a média condicional é uma **parábola**: saber X muda completamente a previsão de Y, e ainda assim a correlação não registra nada.

Se a única evidência disponível fosse o coeficiente de correlação, as duas situações seriam indistinguíveis. **É essa a razão de a hipótese de identificação do próximo capítulo ser $\mathbb{E}(u_i \mid X_i) = 0$, e não $\text{Corr}(u_i, X_i) = 0$.** Um erro que se comportasse como o painel esquerdo satisfaria a segunda condição e violaria a primeira — e a estimativa de MQO seria enganosa exatamente onde mais importa.

Um último detalhe, que fecha o argumento da seção sobre dependência na variância:

```
round(tapply(Y, ifelse(abs(X) > 1, "|X| > 1", "|X| <= 1"), var), 3)
```

```
|X| <= 1   |X| > 1
    0.153     0.826
```

A variância condicional de Y também depende de X — é mais de cinco vezes maior nas pontas. Dependência na média e dependência na variância, ambas invisíveis à correlação.

3.12 Exercícios

1. **Binomial e o coeficiente binomial.** Um goleiro defende cada pênalti com probabilidade 0,25, independentemente. Em uma disputa de 5 cobranças: (a) escreva a distribuição de probabilidade do número de defesas; (b) calcule $P(X = 2)$ com e sem o coeficiente binomial e explique, em termos de contagem de sequências, por que a segunda está errada; (c) verifique que as seis probabilidades somam 1.

2. **CDF e monotonicidade.** Demonstre que a distribuição acumulada de uma variável aleatória discreta nunca decresce e que $\lim_{x\to\infty} F(x) = 1$. Em seguida, explique por que, no caso contínuo, $P(a < X < b) = P(a \le X \le b)$, mas no discreto isso é falso em geral. Dê um exemplo numérico da falha.

3. **As duas formas da variância.** Demonstre a identidade $\mathrm{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$ a partir da definição, usando apenas a linearidade do valor esperado. Em seguida, use-a para calcular a variância de uma Bernoulli(θ) e mostre que ela é máxima em $\theta = 1/2$. Interprete economicamente esse resultado.

4. **Chebyshev, na mão.** (a) Use a desigualdade de Chebyshev para encontrar o menor k tal que ao menos 95% da massa esteja em $[\mu - k\sigma, \mu + k\sigma]$, para qualquer distribuição. (b) Compare com o k correspondente sob normalidade. (c) Construa uma variável aleatória discreta de três pontos em que o limite de Chebyshev para $k = 2$ seja atingido com igualdade ou quase — por que essas distribuições "extremais" são necessariamente degeneradas?

5. **Padronização e transformações.** Seja X com $\mu_X = 10$ e $\sigma_X = 4$. (a) Obtenha média e variância de $W = 3 - 2X$. (b) Verifique que $Z = (X - \mu_X)/\sigma_X$ tem média 0 e variância 1. (c) Mostre que a correlação entre X e W é exatamente -1, e explique por que isso não depende dos valores de μ_X e σ_X.

6. **Tabela chuva × commute.** Usando a Tabela 1: (a) calcule $P(\text{Rain} \mid \text{Long C.})$ e explique por que ela difere de $P(\text{Long C.} \mid \text{Rain})$;

   (b) obtenha $\mathbb{E}(X \mid Y = 0)$ e $\mathbb{E}(X \mid Y = 1)$; (c) verifique numericamente a lei das expectativas iteradas na direção oposta à do laboratório, isto é, recuperando $\mathbb{E}(X)$ a partir de $\mathbb{E}(X \mid Y)$; (d) que valor deveria ter a célula (Long, Rain) para que X e Y fossem independentes, mantidas as marginais?

7. **Lei das expectativas iteradas e o paradoxo de Simpson.** Uma empresa tem dois departamentos. No departamento A, o salário médio de mulheres é maior que o de homens; no B, também. Mostre, construindo um exemplo numérico com quatro grupos, que ainda assim o salário médio geral das mulheres pode ser **menor** que o dos homens. (a) Use a lei das expectativas iteradas para identificar exatamente o que produz a reversão. (b) O que esse exemplo diz sobre agregar médias sem os pesos corretos?

8. **Covariância zero e a hipótese do próximo capítulo.** Seja X com distribuição simétrica em torno de zero e $Y = X^2$. (a) Prove que $\mathrm{Cov}(X, Y) = 0$. (b) Calcule $\mathbb{E}(Y \mid X)$ e mostre que depende de X. (c) Agora considere $Y = X\varepsilon$, com $\varepsilon \perp X$, $\mathbb{E}(\varepsilon) = 0$: mostre que $\mathbb{E}(Y \mid X) = \mathbb{E}(Y)$ mas que X e Y **não** são independentes. (d) Ordene as três condições — independência, $\mathbb{E}(Y \mid X) = \mathbb{E}(Y)$ e $\mathrm{Corr}(X, Y) = 0$ — da mais forte para a mais fraca, e explique por que a econometria escolhe a **do meio** como hipótese de identificação em

   1.

4 Regressão Linear com um Regressor

Os dois capítulos anteriores prepararam o terreno sem ainda estimar nada. O Seção 2 estabeleceu *o que* queremos — efeitos causais — e por que dados observacionais tornam a obtenção deles difícil. O Seção 3 recuperou as ferramentas, e uma delas em particular: a **média condicional** $\mathbb{E}(Y \mid X)$, apresentada ali como a melhor previsão de Y dado que conhecemos X. Este capítulo une as duas coisas. Ele apresenta o **modelo de regressão linear com um regressor** e o estimador de **mínimos quadrados ordinários**, e a tese que organiza tudo o que vem a seguir é simples de enunciar e vale a pena fixar desde já:

> A regressão linear é um método para estimar médias condicionais. Toda a dificuldade da inferência causal está concentrada na hipótese que permite interpretar essa média condicional como um efeito causal.

O capítulo corresponde ao Capítulo 4 de Stock e Watson (2020) e à Aula 3 da disciplina. Sua estrutura segue a do professor: primeiro o modelo e o que cada símbolo significa; depois a estimação por MQO, derivada do princípio de minimização até a fórmula fechada; em seguida as medidas de aderência, com o alerta — que merece destaque — de que R^2 alto não é evidência de causalidade; e por fim as **três hipóteses de MQO** e a distribuição amostral dos estimadores. A primeira dessas hipóteses, $\mathbb{E}(u_i \mid X_i) = 0$, é o núcleo conceitual não só do capítulo mas do curso inteiro.

4.1 O modelo de regressão linear

A pergunta empírica que atravessa o capítulo é a que abre o Capítulo 4 de Stock e Watson (2020): **uma redução na razão aluno-professor melhora o desempenho acadêmico?** É uma pergunta de política pública com consequências orçamentárias imediatas — contratar professores custa caro, e vale saber quanto se compra com esse dinheiro. E é uma pergunta *quantitativa*, no sentido preciso do Seção 1: não basta saber que turmas menores "ajudam"; é preciso saber de quanto.

Os dados são de 420 distritos escolares da Califórnia. Para cada distrito observamos o desempenho médio dos alunos em um teste padronizado ($TestScore$) e a razão aluno-professor ($ClassSize$, ou STR, de *student–teacher ratio*). O problema é estimar uma equação simples:

$$TestScore = \beta_0 + \beta_1 \cdot ClassSize + \text{erro}$$

O objeto de interesse é β_1, e sua definição é a de uma derivada — a variação em $TestScore$ associada a uma variação unitária em $ClassSize$:

$$\boxed{\beta_1 = \frac{\Delta TestScore}{\Delta ClassSize}}$$

Não sabemos o verdadeiro valor de β_1. Ele é uma característica da população — fixo, determinado, desconhecido —, e precisamos **estimá-lo** a partir dos dados observacionais de que dispomos. É exatamente a distinção parâmetro *versus* estatística que as Notas de Estatística assentam no capítulo de estimação pontual, transplantada agora para um contexto em que o parâmetro não é uma média, mas a inclinação de uma reta.

Generalizando, considere uma amostra de tamanho n:

$$\boxed{Y_i = \beta_0 + \beta_1 X_i + u_i, \qquad i = 1, 2, \ldots, n}$$

Esta é a equação central do capítulo, e cada símbolo tem nome:

Símbolo	Nome	O que é
Y_i	variável dependente	o que queremos explicar (o *regressando*)
X_i	variável independente ou **regressor**	o que usamos para explicar
β_0	**intercepto**	valor de $\mathbb{E}(Y \mid X)$ quando $X = 0$
β_1	**coeficiente angular**	efeito de uma unidade a mais de X sobre $\mathbb{E}(Y \mid X)$
u_i	**erro** da regressão	tudo o mais que afeta Y e não é X

A equação recebe o nome de **modelo de regressão linear com um regressor** — "com um regressor" porque há uma única variável explicativa do lado direito, além do intercepto. O 1 é, nesse sentido, o caso mais simples possível, e é justamente por isso que ele é o lugar certo para entender o que está em jogo: tudo o que o curso fará depois — regressão múltipla, painel, variáveis instrumentais — é uma elaboração sobre esta estrutura.

4.1.1 O termo de erro

Por que introduzir u_i? Há duas razões, e o professor as apresenta juntas porque de fato operam juntas.

A primeira é de **forma funcional**. Estamos aproximando a relação entre Y e X por uma função linear. Não há nenhuma garantia de que a verdadeira relação seja uma reta; a aproximação certamente contém erro, e esse erro precisa ir para algum lugar.

A segunda, e mais importante, é de **omissão**. Com certeza existem outros fatores determinantes do desempenho acadêmico além do tamanho da turma: a qualidade dos professores, a infraestrutura da escola, a renda das famílias, o envolvimento dos pais, a fração de alunos que não fala inglês em casa. Nenhum deles está do lado direito da equação. É plausível, portanto, introduzir um termo que **represente os fatores não observados que afetam** Y.

O erro é, assim, um contêiner: ele carrega tudo o que influencia Y_i e não é X_i. Essa leitura será decisiva mais adiante — quando enunciarmos a hipótese $\mathbb{E}(u_i \mid X_i) = 0$, o que estaremos exigindo é uma propriedade *desse contêiner*, e é aí que a inferência causal se ganha ou se perde.

4.1.2 A Função de Regressão Populacional

Tome a esperança condicional a X dos dois lados de $Y_i = \beta_0 + \beta_1 X_i + u_i$. Pela linearidade da esperança,

$$\mathbb{E}(Y \mid X) = \beta_0 + \beta_1 X + \mathbb{E}(u \mid X).$$

Se **assumirmos** que $\mathbb{E}(u \mid X) = 0$ — a hipótese de média condicional igual a zero —, o último termo desaparece e resta

$$\boxed{\mathbb{E}(Y \mid X) = \beta_0 + \beta_1 X}$$

Esta equação chama-se **Função de Regressão Populacional** (*Population Regression Function*, PRF). Ela é a peça que conecta este capítulo ao anterior, e vale gastar um parágrafo nela.

> **i** Ferramenta estatística — a média condicional
>
> A PRF **é** uma média condicional. Não é uma analogia: $\mathbb{E}(Y \mid X)$ é exatamente o objeto definido no Seção 3 — a média da distribuição de Y entre os elementos da população que têm aquele valor de X —, e o que a PRF acrescenta é uma **hipótese sobre sua forma**: a de que, como função de X, ela é uma reta de intercepto β_0 e inclinação β_1.
> Nas **Notas de Aula de Estatística** (Prof. Eduardo Campos), a média condicional aparece no tratamento de distribuições conjuntas, com a lei das esperanças iteradas $\mathbb{E}(Y) = \mathbb{E}[\mathbb{E}(Y \mid X)]$ e a interpretação da esperança condicional como melhor previsão de Y dado X no sentido do erro quadrático médio.
> É por isso que aquele capítulo veio antes deste. Quando se diz que "a regressão estima médias condicionais", isso deve ser lido literalmente: estimar β_0 e β_1 é estimar a função $x \mapsto \mathbb{E}(Y \mid X = x)$.

Com a PRF em mãos, o termo de erro ganha uma definição precisa em vez de uma descrição vaga. Ele é a **diferença entre Y_i e o valor predito de Y_i pela PRF**. No exemplo do capítulo:

$$\begin{aligned} TestScore &= \beta_0 + \beta_1\, ClassSize + \text{erro}, \\ \mathbb{E}(TestScore \mid ClassSize) &= \beta_0 + \beta_1\, ClassSize, \\ \text{erro} &= TestScore - \mathbb{E}(TestScore \mid ClassSize). \end{aligned}$$

O erro é o desvio da observação individual em relação à média de seu grupo. Um distrito com erro positivo teve desempenho acima do que se esperaria de distritos com aquele mesmo tamanho de turma; um distrito com erro negativo, abaixo. E a hipótese $\mathbb{E}(u \mid X) = 0$ diz que esses desvios, dentro de cada grupo definido por um valor de X, se cancelam em média — o que é uma tautologia se $\mathbb{E}(Y \mid X)$ é *de fato* linear, e uma hipótese com conteúdo empírico quando lida, como faremos na 1, como afirmação sobre os fatores omitidos.

! Uma distinção que será cobrada o curso inteiro

$\mathbb{E}(u \mid X) = 0$ é **mais forte** que $\mathbb{E}(u) = 0$. A segunda condição diz apenas que os fatores omitidos se cancelam *na população como um todo* — e é, na verdade, uma normalização sem conteúdo, já que qualquer média não nula do erro pode ser absorvida pelo intercepto. A primeira diz que eles se cancelam **dentro de cada valor de** X, o que é uma afirmação sobre a *relação* entre os fatores omitidos e o regressor. É essa a hipótese que dá conteúdo causal a β_1.

4.2 Estimando os coeficientes: mínimos quadrados ordinários

A PRF é populacional: β_0 e β_1 existem, são fixos e não os conhecemos. Temos uma amostra de n pares (X_i, Y_i) e queremos, a partir dela, produzir estimativas. O critério que usaremos é o de **mínimos quadrados ordinários** (MQO, ou OLS).

A ideia é natural. Cada par candidato (b_0, b_1) define uma reta, e cada reta comete, em cada observação, um erro de previsão $Y_i - b_0 - b_1 X_i$. Queremos a reta que torne esses erros pequenos no conjunto. Elevá-los ao quadrado — em vez de tomar o valor absoluto — penaliza desvios grandes mais que proporcionalmente, e tem a vantagem decisiva de produzir um problema com solução analítica. O estimador de MQO é a solução de

$$\boxed{\min_{b_0,\, b_1} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2}$$

4.2.1 As condições de primeira ordem

O problema é de minimização irrestrita de uma função quadrática e estritamente convexa em (b_0, b_1); as condições de primeira ordem são necessárias e suficientes. Seja $S(b_0, b_1) = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$. Derivando em relação a b_0, pela regra da cadeia,

$$\frac{\partial S}{\partial b_0} = \sum_{i=1}^{n} 2\,(Y_i - b_0 - b_1 X_i) \cdot (-1) = -2 \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) = 0,$$

e em relação a b_1 — que é a derivada que o professor escreve explicitamente no slide 26 —,

$$\frac{\partial S}{\partial b_1} = \frac{\partial \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2}{\partial b_1} = -2 \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)\, X_i = 0.$$

Dividindo ambas por -2 e denotando por $\hat{\beta}_0, \hat{\beta}_1$ a solução, obtemos o sistema que define o estimador de MQO. Como $\hat{u}_i \equiv Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$ é o **resíduo** da observação i, o sistema tem uma leitura imediata:

! Condições de ortogonalidade

$$\begin{array}{lll} \text{(i)} & \displaystyle\sum_{i=1}^{n} \hat{u}_i = 0 & \iff \displaystyle\sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) = 0 \\ \text{(ii)} & \displaystyle\sum_{i=1}^{n} \hat{u}_i X_i = 0 & \iff \displaystyle\sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) X_i = 0 \end{array}$$

Os resíduos de MQO somam zero e são ortogonais ao regressor. Estas duas identidades valem **por construção**, em qualquer amostra, quaisquer que sejam os dados — não são hipóteses, não são resultados assintóticos, não podem ser testadas. São a definição do estimador escrita de outra forma.

A distinção entre as condições de ortogonalidade (amostrais, sempre válidas) e a hipótese $\mathbb{E}(u_i \mid X_i) = 0$ (populacional, possivelmente falsa) é uma das que mais confundem quem começa. O MQO *sempre* produz resíduos ortogonais a X; isso não diz absolutamente nada sobre se o **erro** populacional é ortogonal a X. A primeira é uma propriedade aritmética da conta que fizemos; a segunda é uma afirmação sobre o mundo.

4.2.2 Resolvendo o sistema

Vamos agora à derivação algébrica que leva às fórmulas fechadas, seguindo o encadeamento dos slides 18–20.

Passo 1. Divida a primeira condição por n:

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) = 0 \quad \Longrightarrow \quad \bar{Y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{X} = 0 \quad \Longrightarrow \quad \bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}.$$

Este resultado já é interessante em si: **a reta de MQO passa pelo ponto médio da nuvem de pontos** $(\bar{X}, \bar{Y})$. Ele também nos dá imediatamente $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$, restando determinar $\hat{\beta}_1$.

Passo 2. Substitua $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ na segunda condição:

$$\sum_{i=1}^{n}\left(Y_i - \bar{Y} + \hat{\beta}_1 \bar{X} - \hat{\beta}_1 X_i\right) X_i = 0 \quad \Longrightarrow \quad \sum_{i=1}^{n}\Big[\left(Y_i - \bar{Y}\right) - \hat{\beta}_1\left(X_i - \bar{X}\right)\Big] X_i = 0.$$

Separando os dois termos,

$$\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right) X_i \;=\; \hat{\beta}_1 \sum_{i=1}^{n}\left(X_i - \bar{X}\right) X_i.$$

Passo 3 — o truque de centragem. As duas somas acima estão em uma forma "meio centrada": o primeiro fator é um desvio em relação à média, o segundo é o valor bruto. O professor mostra que isso não faz diferença. Considere:

$$\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2=\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(X_i-\bar{X}\right)=\sum_{i=1}^{n}\left(X_i-\bar{X}\right)X_i\;-\;\bar{X}\underbrace{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)}_{=0},$$

onde o último somatório é nulo porque os desvios em relação à média amostral sempre somam zero — $\sum(X_i-\bar{X})=\sum X_i-n\bar{X}=n\bar{X}-n\bar{X}=0$. Logo:

$$\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2=\sum_{i=1}^{n}\left(X_i-\bar{X}\right)X_i.$$

Pelo mesmo raciocínio, subtraindo $\bar{X}$ do segundo fator do lado esquerdo,

$$\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)X_i=\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)\left(X_i-\bar{X}\right)+\bar{X}\underbrace{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)}_{=0}=\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)\left(X_i-\bar{X}\right).$$

Passo 4. Combinando as três últimas equações, chegamos ao resultado central:

! Os estimadores de MQO

$$\hat{\beta}_1=\frac{\displaystyle\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{\displaystyle\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}\qquad\qquad\hat{\beta}_0=\bar{Y}-\hat{\beta}_1\bar{X}$$

4.2.3 Covariância sobre variância

A fórmula de $\hat{\beta}_1$ tem uma leitura que vale mais que a álgebra. Divida numerador e denominador por $n-1$:

$$\hat{\beta}_1=\frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}=\frac{s_{XY}}{s_X^2}$$

onde s_{XY} é a **covariância amostral** entre X e Y e s_X^2 é a **variância amostral** de X. Ou seja:

$$\hat{\beta}_1=\frac{s_{XY}}{s_X^2}$$

O estimador de MQO é a razão entre a covariância e a variância — e a contrapartida populacional dessa expressão, $\mathrm{Cov}(X,Y)/\mathrm{Var}(X)$, é exatamente o coeficiente da projeção linear de Y em X definida no Seção 3. O MQO substitui momentos populacionais por momentos amostrais. É esse o princípio de analogia que reaparecerá, sob o nome de **método dos momentos**, em praticamente todo estimador do curso.

Três consequências imediatas se leem na fórmula. O **sinal** de $\hat{\beta}_1$ é o sinal da covariância amostral: se X e Y variam juntos, a inclinação é positiva. Se X não varia na amostra, $s_X^2 = 0$ e $\hat{\beta}_1$ **não existe** — não há como estimar o efeito de uma variável que não varia, o que antecipa a exigência de "variação suficiente no regressor". E a magnitude depende das unidades de medida das duas variáveis, o que motiva a versão padronizada que aparecerá no item [c] do exercício da 1.

4.3 Interpretação: o caso TestScore

Aplicado aos 420 distritos californianos, o MQO produz a reta estimada que o professor apresenta no slide 23:

$$\boxed{\widehat{TestScore} = 698{,}9 - 2{,}28 \cdot STR}$$

A interpretação do **coeficiente angular** é a que interessa. $\hat{\beta}_1 = -2{,}28$ significa:

> Distritos com **um aluno por professor a mais** têm, em média, um $TestScore$ mais baixo em 2,28 unidades do teste.

Três cuidados de leitura, que o professor faz questão de separar. Primeiro, "em média": a afirmação é sobre a média condicional, não sobre nenhum distrito individual — há distritos com turmas grandes e desempenho excelente. Segundo, "2,28 unidades do teste": a interpretação de um coeficiente é sempre nas **unidades da variável dependente**, e é preciso saber se 2,28 pontos é muito ou pouco na escala do teste (com desvio-padrão de cerca de 19 pontos, é pouco — cerca de 0,12 desvio-padrão). Terceiro, **"distritos com um aluno a mais *têm*", não "reduzir a turma em um aluno *causaria*"**. A segunda leitura só é legítima sob a Hipótese 1, que discutiremos adiante, e há razões fortes para duvidar dela neste exemplo: distritos com turmas menores tendem a ser mais ricos, e a renda familiar também afeta o desempenho.

E o **intercepto**? $\hat{\beta}_0 = 698{,}9$ seria, literalmente, o $TestScore$ médio predito de um distrito com $STR = 0$. O professor pergunta, no slide, por que essa interpretação não tem significado econômico. A resposta tem duas camadas:

1. $STR = 0$ **é economicamente absurdo.** Uma razão aluno-professor igual a zero significa uma escola com alunos e nenhum professor — ou com professores e nenhum aluno. Não é um estado do mundo sobre o qual faça sentido fazer previsões.
2. **E, mais importante,** $STR = 0$ **está muito fora do suporte dos dados.** Na amostra, STR varia aproximadamente entre 14 e 25. A reta foi ajustada nesse intervalo, e nada nos dados informa sobre o comportamento de $\mathbb{E}(Y \mid X)$ em $X = 0$. Avaliar a reta ali é **extrapolação**: estamos estendendo uma aproximação linear local para uma região onde não há observação alguma que a discipline. Mesmo que a relação verdadeira seja bem aproximada por uma reta entre 14 e 25, não há razão para supor que a mesma reta valha em 0.

A função do intercepto, portanto, é essencialmente técnica: ele **posiciona verticalmente a reta** de modo que ela passe por $(\bar{X}, \bar{Y})$, garantindo que os resíduos somem zero. Sem ele, a reta seria forçada pela origem e $\hat{\beta}_1$ seria viesado. Isso não é pouco — mas não é uma interpretação econômica.

4.4 Observações importantes

Antes de seguir, o professor reúne quatro observações que consolidam o que foi feito. Elas parecem pontos de notação e são, na verdade, conceituais.

Parâmetros *versus* estimadores. Os parâmetros populacionais β_0 e β_1 são **fixos e desconhecidos** — números que existem no mundo e que não observamos. Eles são estimados por suas contrapartidas amostrais $\hat{\beta}_0$ e $\hat{\beta}_1$, que são **variáveis aleatórias**: funções da amostra, e portanto diferentes a cada amostra que sorteássemos. É exatamente a distinção estimador/estimativa das Notas de Estatística, e ela é o que torna possível a seção sobre distribuição amostral mais adiante.

Erro *versus* resíduo. O termo de erro $u_i = Y_i - \beta_0 - \beta_1 X_i$ **não é observável**, porque depende dos parâmetros populacionais desconhecidos. Sua contrapartida amostral é o **resíduo** $\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$, que calculamos e usamos como estimativa do erro. A tabela abaixo organiza o paralelismo:

População (não observável)	Amostra (calculável)
$\beta_0,\ \beta_1$	$\hat{\beta}_0,\ \hat{\beta}_1$
$u_i = Y_i - \beta_0 - \beta_1 X_i$	$\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$
$\mathbb{E}(Y \mid X) = \beta_0 + \beta_1 X$	$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
$\mathbb{E}(u \mid X) = 0$ (hipótese, pode falhar)	$\sum \hat{u}_i X_i = 0$ (identidade, sempre vale)

A última linha é a mais importante da tabela, e reencontra o ponto feito acima: a coluna da direita vale sempre; a da esquerda é a aposta.

A decomposição do valor observado. Toda observação se escreve como parte explicada mais parte residual:

$$\boxed{Y_i = \hat{Y}_i + \hat{u}_i, \qquad \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i}$$

Essa decomposição é a base da seção sobre medidas de aderência: o R^2 nada mais é do que a pergunta "que fração da variação de Y está no primeiro pedaço?".

A leitura geométrica. As condições de ortogonalidade têm uma interpretação em álgebra linear que o professor menciona e que vale desenvolver, porque é ela que dá o nome "ortogonalidade". Pense nas n observações como vetores em $\mathbb{R}^n$: o vetor $\mathbf{Y} = (Y_1, \dots, Y_n)'$, o vetor de uns $\mathbf{1} = (1, \dots, 1)'$ e o vetor $\mathbf{X} = (X_1, \dots, X_n)'$. As duas condições dizem que o vetor de resíduos $\hat{\mathbf{u}}$ tem **produto interno nulo** com ambos:

$$\hat{\mathbf{u}}'\mathbf{1} = \sum_{i=1}^{n} \hat{u}_i = 0, \qquad \hat{\mathbf{u}}'\mathbf{X} = \sum_{i=1}^{n} \hat{u}_i X_i = 0.$$

Ou seja, $\hat{\mathbf{u}}$ é **ortogonal ao plano gerado pelos regressores** (o subespaço de $\mathbb{R}^n$ gerado por $\mathbf{1}$ e $\mathbf{X}$). E o vetor de valores ajustados $\hat{\mathbf{Y}} = \mathbf{Y} - \hat{\mathbf{u}}$ é a **projeção ortogonal** de $\mathbf{Y}$ sobre esse plano. Minimizar a soma de quadrados dos resíduos é minimizar $\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2$, a distância euclidiana entre $\mathbf{Y}$ e o plano — e a solução de um problema de distância mínima a um subespaço é, sempre, a projeção ortogonal. O MQO *é* uma projeção.

Na notação geral do professor, com K regressores mais o intercepto ($x_0 \equiv 1$),

$$\sum_{i=1}^{n} \hat{u}_i X_{ki} = 0, \qquad k = 0, 1, 2, \ldots, K.$$

Este é o ponto de partida do tratamento matricial da regressão múltipla, e a razão pela qual $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ tem a forma que tem. Por ora, basta reter a imagem: **ajustar por MQO é projetar Y no espaço dos regressores, e o resíduo é o que sobra, perpendicular a tudo o que foi usado para explicar.**

4.5 Exemplo: o "beta" de uma ação

O professor abre um parêntese que merece atenção especial de um público de Economia e Finanças. O celebrado **Capital Asset Pricing Model** (CAPM) relaciona o excesso de retorno esperado de um ativo com o excesso de retorno esperado de um portfólio amplo, o portfólio de mercado:

$$\boxed{R_i - R_f = \beta_i \left(R_m - R_f\right)} \qquad \text{com} \qquad \boxed{\beta_i = \frac{\mathrm{Cov}(R_i, R_m)}{\mathrm{Var}(R_m)}}$$

onde R_i é o retorno do ativo i, R_m o retorno do portfólio de mercado e R_f a taxa livre de risco.

Compare a segunda expressão com a fórmula de MQO obtida acima, $\hat{\beta}_1 = s_{XY}/s_X^2$. **É a mesma fórmula** — a versão populacional da versão amostral. A conclusão é forte e vale enunciar:

> **!** O beta de uma ação é um coeficiente de regressão
>
> O β de um ativo, tal como definido pelo CAPM e reportado por qualquer terminal financeiro, **é literalmente o coeficiente angular da regressão do excesso de retorno do ativo sobre o excesso de retorno do mercado**:
>
> $$\left(R_{it} - R_{ft}\right) = \alpha_i + \beta_i \left(R_{mt} - R_{ft}\right) + u_{it}.$$
>
> Estimar o beta de uma ação é rodar uma regressão simples. Nada mais.

Isso ilumina o CAPM e a regressão simultaneamente.

Do lado do CAPM, explica-se de onde vem a interpretação usual: β_i mede a sensibilidade do ativo aos movimentos do mercado. Um ativo com $\beta = 1{,}5$ tende a mover-se 1,5% quando o mercado se move 1% — é a leitura de coeficiente angular, aplicada a retornos. Ativos com $\beta > 1$ são "agressivos"; com $0 < \beta < 1$, "defensivos"; com $\beta < 0$, funcionam como *hedge*. E o CAPM prediz $\alpha_i = 0$: o intercepto da regressão deveria ser nulo, de modo que testar a validade do CAPM para um ativo é testar $H_0 : \alpha_i = 0$ — um teste de hipótese sobre um coeficiente de regressão, exatamente o que a Aula 4 nos dará ferramentas para fazer.

Do lado da regressão, o exemplo mostra que a fórmula covariância-sobre-variância não é uma curiosidade algébrica: ela é o objeto que a teoria financeira identificou, de forma independente, como a medida relevante de risco sistemático. O CAPM decompõe o risco total de um ativo em uma parte que covaria com o mercado — capturada por β_i, não diversificável e por isso remunerada em

equilíbrio — e uma parte idiossincrática, que é justamente o resíduo u_{it} da regressão, diversificável e não remunerada. **A decomposição $Y_i = \hat{Y}_i + \hat{u}_i$ é, no CAPM, a decomposição do retorno em componente sistemático e componente idiossincrático.** Que ela apareça em um curso de econometria e em um curso de finanças com o mesmo conteúdo matemático não é coincidência: é a mesma projeção ortogonal.

4.6 Medidas de aderência

Uma vez estimada a reta de regressão, é natural perguntar se ela é uma boa descrição dos dados. Duas estatísticas respondem, de maneiras complementares: o **coeficiente de determinação** R^2, que é adimensional e mede a *fração* da variação explicada, e o **erro-padrão da regressão** (SER), que é medido nas unidades de Y e mede a *magnitude* típica do erro.

4.6.1 As três somas de quadrados

Partimos da decomposição $Y_i = \hat{Y}_i + \hat{u}_i$ e definimos:

$$\boxed{\begin{array}{ll} \text{TSS} = \displaystyle\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 & \text{(Soma Total dos Quadrados)} \\ \text{ESS} = \displaystyle\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 & \text{(Soma dos Quadrados Explicados)} \\ \text{SSR} = \displaystyle\sum_{i=1}^{n} \hat{u}_i^2 & \text{(Soma dos Quadrados dos Resíduos)} \end{array}}$$

A TSS mede a variação total de Y na amostra — é $(n-1)s_Y^2$. A ESS mede a variação dos **valores preditos** em torno da média. A SSR mede a variação que ficou no resíduo.

Note que na definição de ESS os valores preditos são comparados a $\bar{Y}$, a média dos valores **observados**. O professor pergunta, no slide 29: por que se pode fazer isso? Porque $\bar{\hat{Y}} = \bar{Y}$, isto é, **a média dos valores ajustados é igual à média dos valores observados**. A demonstração é uma linha e usa a primeira condição de ortogonalidade:

$$\frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \hat{u}_i) = \bar{Y} - \frac{1}{n}\underbrace{\sum_{i=1}^{n} \hat{u}_i}_{=0} = \bar{Y}.$$

É o mesmo fato geométrico de que a reta de MQO passa por $(\bar{X}, \bar{Y})$, visto de outro ângulo: como $\hat{Y}_i$ é uma função afim de X_i, sua média é $\hat{\beta}_0 + \hat{\beta}_1 \bar{X}$, que é precisamente $\bar{Y}$. Vale notar que essa propriedade depende da presença do intercepto — em uma regressão sem intercepto, $\sum \hat{u}_i \neq 0$ em geral, e boa parte do que vem a seguir deixa de valer.

4.6.2 A decomposição TSS = ESS + SSR

O professor deixa esta demonstração como exercício. Vamos fazê-la, porque ela é o que justifica a interpretação do R^2 como "fração explicada" — sem ela, ESS/TSS seria uma razão qualquer, não uma proporção.

! Proposição — decomposição da variação

$$\boxed{\text{TSS} = \text{ESS} + \text{SSR}}$$

Demonstração. Some e subtraia $\hat{Y}_i$ dentro do quadrado da TSS:

$$\text{TSS} = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n} \Big[\underbrace{\left(Y_i - \hat{Y}_i\right)}_{=\hat{u}_i} + \left(\hat{Y}_i - \bar{Y}\right) \Big]^2 .$$

Expandindo o quadrado do binômio:

$$\text{TSS} = \underbrace{\sum_{i=1}^{n} \hat{u}_i^2}_{=\text{SSR}} + \underbrace{\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2}_{=\text{ESS}} + 2 \underbrace{\sum_{i=1}^{n} \hat{u}_i \left(\hat{Y}_i - \bar{Y}\right)}_{=C} .$$

Resta mostrar que o termo cruzado C é nulo. Substituindo $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$:

$$C = \sum_{i=1}^{n} \hat{u}_i \left(\hat{\beta}_0 + \hat{\beta}_1 X_i - \bar{Y}\right) = \left(\hat{\beta}_0 - \bar{Y}\right) \underbrace{\sum_{i=1}^{n} \hat{u}_i}_{=0} + \hat{\beta}_1 \underbrace{\sum_{i=1}^{n} \hat{u}_i X_i}_{=0} = 0,$$

onde as duas somas se anulam pelas condições de ortogonalidade (i) e (ii), respectivamente. Logo $\text{TSS} = \text{ESS} + \text{SSR}$. ■

Duas observações sobre a demonstração. Primeira: ela usa **as duas** condições de ortogonalidade, e só elas — nenhuma hipótese sobre o processo gerador dos dados. A decomposição é, portanto, uma identidade algébrica que vale em qualquer amostra, mesmo que o modelo esteja completamente errado. Segunda: geometricamente, $\text{TSS} = \text{ESS} + \text{SSR}$ é o **teorema de Pitágoras** aplicado ao triângulo retângulo formado pelos vetores $(\mathbf{Y} - \bar{Y}\mathbf{1})$, $(\hat{\mathbf{Y}} - \bar{Y}\mathbf{1})$ e $\hat{\mathbf{u}}$ — o ângulo reto é exatamente a ortogonalidade que acabamos de usar.

4.6.3 O coeficiente de determinação

! Definição — o R^2

O R^2 da regressão é a **fração da variância amostral de Y que é explicada (ou predita) por X**:

$$\boxed{R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{SSR}}{\text{TSS}}}$$

A segunda igualdade segue de TSS = ESS + SSR: dividindo por TSS, obtém-se $1 = R^2 + \text{SSR/TSS}$.

Propriedades. O R^2 assume valores entre 0 e 1 — o que decorre imediatamente da decomposição, já que ESS e SSR são somas de quadrados, logo não negativas, e ambas não excedem TSS. Os dois casos extremos são instrutivos:

- $R^2 = 0$**.** Se $\hat{\beta}_1 = 0$, então X_i não explica nada da variação de Y_i, e o valor predito é constante: $\hat{Y}_i = \hat{\beta}_0 = \bar{Y}$ para todo i. Nesse caso ESS = 0, os resíduos coincidem com os desvios em relação à média, SSR = TSS, e $R^2 = 0$. A "melhor" previsão que X oferece é a mesma que se faria sem X: a média.
- $R^2 = 1$**.** Se X_i explica *toda* a variação de Y_i, então $\hat{Y}_i = Y_i$ para todo i, todos os resíduos são nulos ($\hat{u}_i = 0\ \forall i$), SSR = 0 e ESS = TSS. Todos os pontos estão exatamente sobre a reta.

Em geral, o R^2 não assume os valores extremos. Um R^2 próximo de 1 indica que o regressor é um **bom preditor** de Y_i; próximo de zero, que não é. E aqui vem o alerta que o professor destaca em negrito no slide 32, e que estas notas destacam ainda mais:

! ALERTA — R^2 alto NÃO é evidência de causalidade

Um R^2 alto não é evidência de causalidade. Ser um bom preditor é muito diferente de ser causal.
As duas coisas são logicamente independentes, e nas duas direções:

- R^2 **alto sem causalidade.** O número de bombeiros enviados a um incêndio prevê extraordinariamente bem o valor dos danos — e não os causa. Vendas de sorvete preveem afogamentos. Em séries de tempo, duas variáveis com tendência prevêem uma à outra com R^2 acima de 0,9 sem qualquer relação. Nada nas fórmulas de ESS, SSR e TSS envolve a hipótese $\mathbb{E}(u \mid X) = 0$: o R^2 é calculável e alto mesmo quando o coeficiente é completamente viesado.
- **Causalidade com R^2 baixo.** No próprio exemplo do capítulo, o R^2 é de apenas 0,051 — o tamanho da turma explica 5% da variação do desempenho entre distritos. Isso **não** significa que o efeito seja nulo ou desinteressante: significa que há muitos outros determinantes do desempenho. Um efeito causal pode ser real, robusto, e economicamente relevante, e ainda assim responder por uma fração pequena da variação total.

A confusão tem uma origem clara. O R^2 é uma medida de **ajuste**, e ajuste é uma propriedade da relação *estatística* entre as variáveis na amostra. Causalidade é uma propriedade da relação *estrutural* entre elas no mundo. A primeira se lê nos dados; a segunda depende de uma hipótese de identificação que os dados não verificam. Maximizar o R^2 é o objetivo correto em um problema de **previsão**; em um problema de **inferência causal**, é um critério irrelevante e potencialmente enganoso — acrescentar regressores sempre aumenta o R^2, inclusive os que introduzem viés.

4.6.4 O exercício do slide 33

O professor propõe três resultados sobre o R^2 e a correlação amostral. Vamos resolvê-los, porque eles amarram tudo o que foi feito nesta seção. Recorde a notação: $r_{XY} = s_{XY}/(s_X s_Y)$ é a correlação amostral, e s_X, s_Y os desvios-padrão amostrais.

[a] Mostre que $R^2 = r_{XY}^2$ na regressão simples.

Comece pela ESS. Como $\hat{Y}_i - \bar{Y} = (\hat{\beta}_0 + \hat{\beta}_1 X_i) - (\hat{\beta}_0 + \hat{\beta}_1 \bar{X}) = \hat{\beta}_1 (X_i - \bar{X})$ — usando $\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}$ do Passo 1 —, temos

$$\text{ESS} = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 = \hat{\beta}_1^2 \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 .$$

Substituindo $\hat{\beta}_1 = s_{XY}/s_X^2$ e escrevendo as somas em termos das variâncias amostrais, $\sum (X_i - \bar{X})^2 = (n-1)s_X^2$ e $\text{TSS} = (n-1)s_Y^2$:

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{\left(\dfrac{s_{XY}}{s_X^2}\right)^2 (n-1) s_X^2}{(n-1) s_Y^2} = \frac{s_{XY}^2}{s_X^4} \cdot \frac{s_X^2}{s_Y^2} = \frac{s_{XY}^2}{s_X^2 s_Y^2} = \left(\frac{s_{XY}}{s_X s_Y}\right)^2 = r_{XY}^2 . \qquad \blacksquare$$

O resultado justifica a notação: o "R" de R^2 é a correlação. **Atenção:** ele vale apenas na regressão **simples**. Com mais de um regressor, R^2 é o quadrado da correlação entre Y e $\hat{Y}$ — a correlação múltipla —, não de nenhuma correlação simples.

[b] Mostre que o R^2 da regressão de Y em X é o mesmo da regressão de X em Y.

Segue imediatamente de [a]. Pelo item anterior, o R^2 da regressão de Y em X é r_{XY}^2. Aplicando o mesmo resultado à regressão de X em Y — que é a mesma conta com os papéis trocados —, seu R^2 é r_{YX}^2. Mas a correlação amostral é **simétrica**, $r_{XY} = s_{XY}/(s_X s_Y) = s_{YX}/(s_Y s_X) = r_{YX}$, pois a covariância é simétrica. Logo os dois R^2 coincidem. $\blacksquare$

Este resultado é conceitualmente importante e frequentemente mal compreendido. **Os coeficientes das duas regressões são diferentes** — a regressão de Y em X tem inclinação s_{XY}/s_X^2 e a de X em Y tem s_{XY}/s_Y^2, e as duas retas não coincidem no plano (exceto quando $|r_{XY}| = 1$). Mas o **ajuste** é o mesmo. A moral: **o R^2 não sabe qual é a variável dependente**. Ele é uma medida de associação, e associação é simétrica. Nenhuma estatística de ajuste pode, portanto, dizer qual variável causa qual — o que é apenas outra forma de enunciar o alerta da caixa anterior.

[c] Mostre que $\hat{\beta}_1 = r_{XY}\,(s_Y/s_X)$.

Direto da definição de correlação. Como $s_{XY} = r_{XY}\, s_X s_Y$,

$$\hat{\beta}_1 = \frac{s_{XY}}{s_X^2} = \frac{r_{XY}\, s_X s_Y}{s_X^2} = r_{XY}\, \frac{s_Y}{s_X} . \qquad \blacksquare$$

A fórmula separa duas coisas que a inclinação mistura. O fator r_{XY}, adimensional e entre -1 e 1, é a **força da associação linear**. O fator s_Y/s_X é um puro **fator de conversão de unidades**: transforma "desvios-padrão de X" em "desvios-padrão de Y". Uma consequência prática: se padronizarmos ambas as variáveis (média zero, desvio-padrão um), então $s_Y = s_X = 1$ e o coeficiente

de MQO **é** a correlação. É por isso que coeficientes de variáveis padronizadas — os *beta coefficients* de alguns softwares — são comparáveis entre si, e coeficientes brutos não são.

Uma segunda consequência é a **regressão à média**, que dá nome ao método. Se X e Y têm o mesmo desvio-padrão (alturas de pais e filhos, no exemplo original de Galton), então $\hat{\beta}_1 = r_{XY}$, que é menor que 1 em valor absoluto sempre que a associação não é perfeita. Filhos de pais muito altos tendem a ser altos, mas **menos** altos que os pais — não por nenhum mecanismo biológico misterioso, mas por uma propriedade aritmética do coeficiente de regressão.

4.6.5 O erro-padrão da regressão

O R^2 é adimensional e responde "que fração?". A segunda medida de aderência responde "de quanto?", nas unidades de Y.

! Definição — erro-padrão da regressão (SER)

O **erro-padrão da regressão** (*Standard Error of the Regression*) é um estimador do desvio-padrão do erro de regressão u_i:

$$\text{SER} = s_{\hat{u}} = \sqrt{s^2_{\hat{u}}}, \qquad s^2_{\hat{u}} = \frac{1}{n-2}\sum_{i=1}^{n} \hat{u}_i^{\,2} = \frac{\text{SSR}}{n-2}$$

onde $\hat{u}_i = Y_i - \hat{Y}_i$.

Por que $n-2$? Pela mesma razão que a variância amostral usa $n-1$: correção de graus de liberdade. Os resíduos não são livres — eles satisfazem as **duas** condições de ortogonalidade, $\sum \hat{u}_i = 0$ e $\sum \hat{u}_i X_i = 0$. Conhecidos $n-2$ resíduos, os dois restantes ficam determinados por essas restrições. Há, portanto, apenas $n-2$ resíduos "livres", e dividir por $n-2$ é o que torna $s^2_{\hat{u}}$ não viesado para σ^2_u sob homocedasticidade.

i Ferramenta estatística — graus de liberdade e o divisor $n-1$

Este é exatamente o argumento que as **Notas de Aula de Estatística** (Prof. Eduardo Campos) desenvolvem para justificar o divisor $n-1$ da variância amostral s^2_Y. Lá, a restrição é uma só — $\sum(Y_i - \bar{Y}) = 0$, decorrente de se ter estimado μ pela própria média amostral —, e a demonstração mostra que $\mathbb{E}\big[\sum(Y_i - \bar{Y})^2\big] = (n-1)\sigma^2$, de modo que dividir por n produziria um estimador viesado para baixo.
A regra geral é: **subtraia do n o número de parâmetros estimados** que entram no cálculo dos desvios. Na variância amostral estima-se um parâmetro (μ), daí $n-1$. Na regressão simples estimam-se dois (β_0 e β_1), daí $n-2$. Na regressão múltipla com k regressores mais intercepto, $n-k-1$. É sempre a mesma contabilidade.
Como em Estatística, a correção importa muito em amostras pequenas e quase nada em amostras grandes: com $n = 420$, $n-2$ e n diferem por menos de meio por cento.

Interpretação. As unidades de $\hat{u}_i$ e de Y_i são as mesmas, de modo que o SER é uma medida da **dispersão das observações em torno da reta de regressão, medida nas unidades da variável dependente**. Se a variável dependente é medida em dólares, o SER mede a magnitude de um desvio típico da reta — a magnitude de um erro de regressão típico — em dólares. No exemplo

do capítulo, SER $\approx$ 18,58 pontos do teste: um distrito típico está a cerca de 19 pontos da reta, o que, comparado ao efeito estimado de 2,28 pontos por aluno, dá a medida exata de quanto do desempenho o tamanho da turma *não* explica.

Comparação com s_Y. Vale pôr as duas fórmulas lado a lado, como faz o professor no slide 37:

$$s_Y = \sqrt{s_Y^2}, \quad s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 \qquad \text{contra} \qquad \text{SER} = \sqrt{s_{\hat{u}}^2}, \quad s_{\hat{u}}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2$$

As fórmulas são parecidas, com duas diferenças: $(Y_i - \bar{Y})$ é substituído por $\hat{u}_i$, e o denominador passa de $n-1$ para $n-2$. A substituição tem significado. s_Y mede a dispersão de Y em torno de sua média **incondicional**; o SER mede a dispersão em torno da **média condicional estimada**. A diferença entre os dois é precisamente o que X explica — e, de fato, a relação entre eles é

$$\text{SER}^2 \approx s_Y^2\left(1 - R^2\right),$$

exata a menos do ajuste de graus de liberdade. As duas medidas de aderência, portanto, são duas leituras da mesma decomposição: o R^2 em termos relativos, o SER em termos absolutos.

4.7 As três hipóteses de MQO

Chegamos ao núcleo conceitual do capítulo. Até aqui, o MQO foi tratado como um procedimento puramente algébrico: dados n pares, produza a reta que minimiza a soma de quadrados. Nada do que fizemos exigiu hipótese alguma — as condições de ortogonalidade, a decomposição TSS = ESS + SSR, as fórmulas do R^2 e do SER valem sempre, mesmo que os dados sejam ruído puro.

A pergunta agora é outra e é a pergunta econométrica: **quando $\hat{\beta}_1$ estima o efeito causal β_1?** A resposta depende da natureza dos dados, e o professor a organiza em torno da comparação com o experimento ideal.

Em um experimento aleatorizado, a diferença entre as médias dos grupos de tratamento e de controle,

$$\mathbb{E}(Y \mid X = 1) - \mathbb{E}(Y \mid X = 0),$$

é um **estimador não viesado do efeito causal do tratamento**. A razão, vista no Seção 2, é que a aleatorização garante que os dois grupos sejam comparáveis em tudo o mais: os fatores omitidos têm a mesma distribuição nos dois braços, porque a alocação foi feita por sorteio e não por escolha. Essa lógica se estende ao modelo de regressão linear e ao estimador de MQO — e o que se segue são as três hipóteses matemáticas sob as quais o MQO produz estimativas do efeito causal.

! Hipótese 1 — a distribuição condicional de u_i dado X_i tem média zero

$$\boxed{\mathbb{E}(u_i \mid X_i) = 0}$$

Esta é uma declaração matemática formal sobre os **outros fatores** contidos em u_i. Ela afirma que esses outros fatores não são relacionados com X_i, no sentido de que, dado um valor de X_i, a média da distribuição desses outros fatores é zero.

É a hipótese mais importante do curso.

Interpretação. No exemplo do capítulo, sabemos que $ClassSize$ é apenas um dos determinantes do desempenho acadêmico. Outros fatores — qualidade dos professores, infraestrutura da escola, características socioeconômicas das famílias — também importam, e o termo de erro representa a contribuição de **todos os fatores não observáveis e omitidos**. A hipótese se traduz, então, em

$$\mathbb{E}\left(\text{Outros Fatores} \mid STR\right) = 0 :$$

> Uma vez que X está fixo, os outros fatores que também são determinantes do $TestScore$ têm, em média, impacto zero no $TestScore$.

Dito de outra forma: comparando dois grupos de distritos que diferem no tamanho da turma, os demais determinantes do desempenho estão **equilibrados** entre os grupos. Essa é precisamente a propriedade que a aleatorização garante em um experimento.

O problema dos dados observacionais. Com dados observacionais, X **não** é alocado aleatoriamente. Distritos escolheram seus tamanhos de turma, e essas escolhas refletem orçamentos, prioridades políticas e a composição socioeconômica local. O melhor que podemos esperar é que X seja **"como se" tivesse sido atribuído aleatoriamente**, no sentido de que $\mathbb{E}(u_i \mid X_i) = 0$.

E aqui está o desconforto que o Seção 1 anunciava: **essa hipótese não é testável**. Ela envolve u_i, que não é observável — e os resíduos $\hat{u}_i$, que observamos, satisfazem a ortogonalidade *por construção*, de modo que nenhum diagnóstico baseado neles pode informar sobre a hipótese populacional. Saber se ela é válida em um exercício empírico concreto "requer conhecimento específico da aplicação e juízo", como diz o slide. É literalmente a *arte* de que fala a definição de econometria: a parte do trabalho que nenhum teorema faz por nós.

4.7.1 Média condicional zero e correlação: um ponto sutil

O professor dedica dois slides a uma sutileza lógica que vale desenvolver com cuidado, porque ela organiza a maneira como se *discute* a Hipótese 1 na prática.

Implicação 1: $\mathbb{E}(u \mid X) = 0 \implies \text{Corr}(u, X) = 0$**.**

Isso foi visto no Seção 3 e a demonstração é curta. Pela lei das esperanças iteradas, $\mathbb{E}(u) = \mathbb{E}[\mathbb{E}(u \mid X)] = \mathbb{E}[0] = 0$. E, pela mesma lei aplicada ao produto,

$$\mathbb{E}(uX) = \mathbb{E}\big[\mathbb{E}(uX \mid X)\big] = \mathbb{E}\big[X\,\mathbb{E}(u \mid X)\big] = \mathbb{E}[X \cdot 0] = 0,$$

de modo que $\text{Cov}(u, X) = \mathbb{E}(uX) - \mathbb{E}(u)\mathbb{E}(X) = 0$ e, portanto, $\text{Corr}(u, X) = 0$.

Implicação 2: a recíproca NÃO vale. Se $\text{Corr}(u_i, X_i) = 0$, **não** é necessariamente verdade que $\mathbb{E}(u_i \mid X_i) = 0$.

A razão é conceitual: a correlação é uma medida de **associação linear**. Ela detecta dependência que se manifeste na forma de uma tendência de subir junto ou descer junto, e é cega para todo o resto. Uma variável pode depender fortemente de outra e ter correlação nula com ela.

O contraexemplo padrão, apresentado no Seção 3, é a dependência quadrática. Seja $X \sim \mathcal{N}(0,1)$ e $u = X^2 - 1$. Então:

$$\mathbb{E}(u \mid X) = X^2 - 1 \neq 0 \quad \text{(depende de } X\text{)},$$

de modo que a Hipótese 1 é **violada** de maneira grosseira: conhecer X informa completamente sobre u. Mas

$$\text{Cov}(u, X) = \mathbb{E}\big[(X^2 - 1)X\big] = \mathbb{E}(X^3) - \mathbb{E}(X) = 0 - 0 = 0,$$

porque os momentos ímpares da Normal padrão são nulos. Correlação zero, dependência total.

Implicação 3 — a contrapositiva, que é a útil:

$$\boxed{\text{Corr}(u_i, X_i) \neq 0 \implies \mathbb{E}(u_i \mid X_i) \neq 0}$$

Esta é apenas a contrapositiva da Implicação 1, logicamente equivalente a ela. E é ela que torna a discussão prática viável.

! Por que se discute a Hipótese 1 em termos de correlação

As três implicações se combinam assim. Como correlação não nula **implica** média condicional não nula, basta encontrar uma razão para suspeitar de correlação entre X_i e u_i para concluir que a Hipótese 1 é inválida. Por isso é **mais conveniente conduzir a discussão sobre a hipótese de média condicional igual a zero em termos da possível correlação entre X_i e u_i**: é um alvo mais fácil de mirar, e é suficiente para derrubar a hipótese.
Em resumo: **se u_i e X_i são correlacionados, então a hipótese de média condicional igual a zero é inválida** — e o MQO é viesado.
O que **não** se pode fazer é o inverso: argumentar que a hipótese vale porque não se vê correlação. Ausência de correlação é condição necessária, não suficiente. É por isso que a Hipótese 1 é formulada em termos de $\mathbb{E}(u \mid X)$, e não de $\text{Corr}(u, X)$ — a versão forte é a que dá as propriedades desejadas; a versão fraca é a que se usa para argumentar contra.

É por essa via que se organiza o resto do curso. Cada estratégia que virá — controlar por observáveis, efeitos fixos, variáveis instrumentais, quase-experimentos — é uma tentativa de tornar defensável a afirmação de que, condicionalmente ao desenho adotado, X não está correlacionado com o que ficou no erro. E o **viés de variável omitida**, tema imediato da regressão múltipla, é exatamente o cálculo do que acontece quando essa correlação não é zero.

! Hipótese 2 — (X_i, Y_i), $i = 1, \ldots, n$, são i.i.d.

$$\boxed{(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} F_{XY}, \qquad i = 1, \ldots, n}$$

Esta hipótese é uma afirmação sobre **como a amostra é extraída**. Se as observações são obtidas por amostragem aleatória de uma população grande, então os pares (X_i, Y_i) podem ser considerados uma amostra independente e identicamente distribuída.

Note a diferença de natureza em relação à Hipótese 1. A primeira é uma afirmação sobre a **estrutura econômica** do problema — sobre a relação entre o regressor e os fatores omitidos. A segunda é sobre o **desenho amostral** — sobre o procedimento que gerou os dados. Uma é substantiva, a outra é de engenharia estatística. Ambas são necessárias, mas por razões diferentes: a Hipótese 1 dá o **não-viés**; a Hipótese 2 é o que permite invocar a Lei dos Grandes Números e o Teorema Central do Limite, e portanto o que dá a **distribuição amostral** de que a inferência depende.

ℹ Ferramenta estatística — amostra aleatória simples, LGN e TCL

A hipótese i.i.d. é exatamente a definição de **amostra aleatória simples** das Notas de Estatística: variáveis independentes e identicamente distribuídas segundo a distribuição da população. A diferença aqui é que o objeto sorteado não é um número, mas um **par** (X_i, Y_i) — sorteamos distritos, e cada distrito traz consigo seu tamanho de turma e seu desempenho, com toda a dependência que houver **entre** as duas variáveis preservada. A independência é *entre observações*, não *entre variáveis*.
É sob essa hipótese que valem a Lei dos Grandes Números e o Teorema Central do Limite, e são esses dois teoremas que produzirão, na próxima seção, a consistência e a normalidade assintótica de $\hat{\beta}_1$.

Exemplo — a equação de Mincer. Mais adiante no curso estimaremos a famosa equação de Mincer, que relaciona salário e escolaridade:

$$\log(wage_i) = \beta_0 + \beta_1 \, Educ_i + u_i.$$

A amostra consiste de n trabalhadores extraídos da população de trabalhadores do Brasil. Nesse caso, **por construção** a amostra é identicamente distribuída — todos os trabalhadores vêm da mesma população. E, se a extração é aleatória, os pares $(wage_i, Educ_i)$ serão independentes entre si. (Vale notar que a Hipótese 1 é bem mais problemática aqui: a habilidade não observada afeta tanto a escolaridade quanto o salário, o que é o exemplo canônico de viés de variável omitida.)

Quando a Hipótese 2 é plausível? Tudo depende do esquema de amostragem.

- **Plausível:** dados de *survey*, gerados por questionários coletados de um subconjunto escolhido aleatoriamente de uma população. PNAD, PME, Census — em geral, podem ser tratados como i.i.d. (com ressalvas sobre desenhos amostrais complexos, que exigem pesos).
- **Não plausível: dados em séries de tempo**. O PIB deste trimestre não é independente do trimestre anterior; a inflação de hoje carrega a de ontem. A dependência temporal é a regra, não a exceção. Isso não impede a análise de regressão — exige um aparato assintótico diferente (estacionariedade, ergodicidade) e erros-padrão que corrijam a autocorrelação.
- **Também problemática:** dados agrupados, em que observações do mesmo grupo (escola, município, firma) compartilham choques. Daí a prática de agrupar erros-padrão por *cluster*.

❗ Hipótese 3 — valores extremos são raros

$$\boxed{0 < \mathbb{E}(X_i^4) < \infty \quad \text{e} \quad 0 < \mathbb{E}(Y_i^4) < \infty}$$

Formalmente, X_i e Y_i têm **quartos momentos finitos e não nulos**. Informalmente: obser-

vações com valores muito extremos são improváveis.

Esta é uma **hipótese técnica**, no sentido de que elimina a possibilidade de haver uma observação que torne enganosos os resultados da regressão por MQO. A razão é simples e vale a pena entendê-la sem a formalidade: **o estimador de MQO minimiza uma soma quadrática**. Se alguma observação é um valor extremo, o desvio correspondente é **elevado ao quadrado**, e sua contribuição ao critério cresce quadraticamente com a distância. Uma única observação suficientemente distante pode, sozinha, dominar a soma e arrastar a reta na sua direção.

Duas observações. A primeira é que a hipótese quase sempre vale na prática — variáveis econômicas limitadas por construção (notas de teste entre 0 e 800, anos de escolaridade entre 0 e 25) têm todos os momentos finitos automaticamente. Ela é enunciada porque é necessária para a demonstração formal da normalidade assintótica, não porque seja empiricamente delicada.

A segunda é que, embora a hipótese formal raramente falhe, o **problema prático** que ela sinaliza é muito real: em qualquer amostra finita, uma observação atípica — frequentemente um erro de digitação, um código de missing lido como número, uma unidade de medida trocada — pode alterar substancialmente as estimativas. Daí a recomendação, que é de ofício e não de teoria: **sempre inspecione os dados graficamente antes de rodar a regressão**. O laboratório deste capítulo mostra numericamente o estrago que uma única observação pode fazer.

4.8 Distribuição amostral dos estimadores de MQO

Reunidas as três hipóteses, podemos enfim tratar $\hat{\beta}_0$ e $\hat{\beta}_1$ como o que eles são.

Porque os estimadores de MQO são calculados a partir de uma **amostra aleatória**, eles também são **variáveis aleatórias** e têm uma distribuição de probabilidade — a **distribuição amostral**. Ela descreve os valores que $\hat{\beta}_0$ e $\hat{\beta}_1$ podem assumir em diferentes amostras aleatórias da mesma população.

Esse é o ponto que faz a ponte com toda a inferência que virá. Se tivéssemos sorteado outros 420 distritos, teríamos obtido $-2{,}31$ ou $-2{,}19$ em vez de $-2{,}28$. A estimativa é um número; o estimador é uma variável aleatória; e é sobre a distribuição dessa variável aleatória que faremos, na Aula 4, afirmações do tipo "com 95% de confiança, β_1 está entre $-3{,}30$ e $-1{,}26$".

i Ferramenta estatística — distribuição amostral

O conceito é o mesmo que as **Notas de Aula de Estatística** desenvolvem para $\bar{X}$: um estimador é função da amostra, logo variável aleatória, logo tem esperança, variância e distribuição. Lá, o resultado central é que $\bar{X}$ é não viesado para μ, tem variância σ^2/n e é aproximadamente Normal por TCL.
Aqui, a estrutura é idêntica, aplicada a um estimador mais complicado. E de fato $\hat{\beta}_1$ **é** uma média disfarçada: pode-se mostrar que $\hat{\beta}_1 - \beta_1 = \frac{\frac{1}{n}\sum(X_i-\bar{X})u_i}{\frac{1}{n}\sum(X_i-\bar{X})^2}$, uma razão entre duas médias amostrais — e é aplicando LGN ao denominador e TCL ao numerador que se obtêm consistência e normalidade assintótica.

! Propriedades dos estimadores de MQO

Sob as Hipóteses 1–3:

1. **Não-viés:** $\mathbb{E}(\hat{\beta}_0) = \beta_0$ e $\mathbb{E}(\hat{\beta}_1) = \beta_1$.
2. **Consistência:** $\hat{\beta}_1 \xrightarrow{p} \beta_1$, isto é, $\operatorname{plim}_{n\to\infty} \hat{\beta}_1 = \beta_1$.
3. **Normalidade assintótica:** em amostras grandes, pelo Teorema Central do Limite,

$$\hat{\beta}_1 \overset{a}{\sim} \mathcal{N}\left(\beta_1,\, \sigma^2_{\hat{\beta}_1}\right),$$

com

$$\boxed{\sigma^2_{\hat{\beta}_1} = \operatorname{Var}(\hat{\beta}_1) = \frac{1}{n} \cdot \frac{\operatorname{Var}[(X_i - \mu_X)\, u_i]}{[\operatorname{Var}(X_i)]^2}}$$

A fórmula da variância é a de Stock e Watson (2020) (equação 4.21) e merece ser lida com atenção, porque ela diz **o que faz um estimador ser preciso**. Três fatores:

1. O tamanho da amostra n. A variância cai a taxa $1/n$, de modo que o desvio-padrão cai a taxa $1/\sqrt{n}$. Quadruplicar a amostra reduz o erro-padrão pela metade. É a mesma taxa de $\bar{X}$, e é a taxa canônica da estatística.

2. A variação do regressor, $\operatorname{Var}(X_i)$. Ela aparece **ao quadrado no denominador**, o que significa que mais variação em X produz estimativas mais precisas. A intuição é geométrica e é forte: para estimar a inclinação de uma reta, é preciso ter pontos espalhados ao longo do eixo horizontal. Se todos os X_i estão amontoados em torno de um mesmo valor, pequenas perturbações nos Y_i giram a reta violentamente — o "braço de alavanca" é curto. Se estão bem espalhados, a reta fica ancorada. No limite, com $\operatorname{Var}(X) = 0$, a inclinação é indeterminada, o que a fórmula registra como variância infinita.

3. A variância do erro. Ela entra pelo numerador, via $\operatorname{Var}[(X_i - \mu_X)u_i]$. Menos ruído não explicado — erros menores em torno da PRF — significa estimativas mais precisas. Sob homocedasticidade ($\operatorname{Var}(u_i \mid X_i) = \sigma_u^2$ constante), a expressão se simplifica para a forma familiar

$$\operatorname{Var}(\hat{\beta}_1) = \frac{\sigma_u^2}{n \operatorname{Var}(X_i)} \qquad \text{estimada por} \qquad \widehat{\operatorname{Var}}(\hat{\beta}_1) = \frac{\mathrm{SER}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2},$$

que exibe os três fatores de forma transparente: erro-padrão pequeno quer dizer **muitos dados, muito espalhados em X, com pouco ruído em Y**.

A forma geral da caixa, com $\operatorname{Var}[(X_i - \mu_X)u_i]$ no numerador, **não** supõe homocedasticidade — ela permite que a dispersão do erro varie com X. É essa fórmula geral que os **erros-padrão robustos à heterocedasticidade** estimam, e é por isso que a saída do professor traz `robust`:

```
regress testscr str, robust

Regression with robust standard errors                Number of obs =     420
                                                      F(  1,   418) =   19.26
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.0512
```

```
                                                      Root MSE      =  18.581
------------------------------------------------------------------------------
             |               Robust
    testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+----------------------------------------------------------------
         str |  -2.279808   .5194892    -4.39   0.000    -3.300945   -1.258671
       _cons |   698.933   10.36436    67.44   0.000     678.5602    719.3057
------------------------------------------------------------------------------
```

Reconhecemos agora quase tudo nesse output: `Coef.` são $\hat{\beta}_0$ e $\hat{\beta}_1$; `R-squared` é ESS/TSS; `Root MSE` é o SER; `Number of obs` é n. O que ainda não sabemos ler é a coluna `Std. Err.` e o que dela decorre — as estatísticas t, os p-valores, os intervalos de confiança. Esse é precisamente o assunto da **Aula 4**, que tratará de testes de hipóteses e intervalos de confiança na regressão simples, e que fecha o tratamento do Capítulo 4 de Stock e Watson (2020). Por ora, o essencial é que a coluna existe porque $\hat{\beta}_1$ é uma variável aleatória, e que sua magnitude é governada pelos três fatores que acabamos de discutir.

Uma última observação de software. Os slides do curso usam STATA, onde a opção `robust` produz os erros-padrão de White (variante HC1). O laboratório a seguir reproduz exatamente esse cálculo em R, através da função `ep_robusto()` definida no preâmbulo destas notas — implementada a partir da fórmula, e não importada de um pacote, para deixar visível o que o software faz por baixo do capô.

4.9 Laboratório computacional em R

O laboratório deste capítulo é o mais extenso das notas até aqui, e faz sete coisas: reproduz a regressão do exemplo, calcula o MQO "na mão" e confere contra o `lm()`, verifica numericamente as condições de ortogonalidade e a decomposição TSS = ESS + SSR, compara erros-padrão robustos e homocedásticos, e — a figura central — constrói por simulação a distribuição amostral de $\hat{\beta}_1$.

4.9.1 Os dados: uma advertência necessária

! Os dados deste laboratório são SIMULADOS

O conjunto de dados de distritos escolares da Califórnia usado por Stock e Watson (2020) (`caschool`, distribuído com o pacote `AER`) **não está disponível** no ambiente em que estas notas são compiladas. Em vez de omitir o exemplo, **simulo dados calibrados** para reproduzir a *ordem de grandeza* dos resultados do slide 34: $n = 420$, $\hat{\beta}_0 \approx 699$, $\hat{\beta}_1 \approx -2{,}28$, $R^2 \approx 0{,}05$ e SER $\approx 18{,}6$.
Os números que aparecem abaixo, portanto, **não são os dados originais de Stock e Watson** e não devem ser citados como tal. São uma reconstrução didática, e a única coisa que herdam do exemplo real são os momentos: média e desvio-padrão de STR compatíveis com a amostra californiana, e uma variância do erro escolhida para produzir o R^2 observado. Tudo o que o laboratório demonstra — identidades algébricas, comportamento dos estimadores, forma da distribuição amostral — vale para quaisquer dados e não depende dessa calibração.

A calibração é simples de justificar. Sob o modelo, $R^2 \approx \beta_1^2 \mathrm{Var}(X)/[\beta_1^2 \mathrm{Var}(X) + \sigma_u^2]$; resolvendo para σ_u com $\beta_1 = -2{,}28$, $\mathrm{sd}(X) = 1{,}89$ e $R^2 = 0{,}051$, obtém-se $\sigma_u \approx 18{,}55$ — que é, coerentemente, o SER do slide.

```
set.seed(5490)
n_d     <- 420
beta0   <- 698.9      # valores "populacionais" usados na simulacao
beta1   <- -2.28
sigma_u <- 18.55      # calibrado para reproduzir R2 ~ 0,051

STR       <- rnorm(n_d, mean = 19.64, sd = 1.89)   # momentos da amostra da California
u         <- rnorm(n_d, mean = 0,     sd = sigma_u)
TestScore <- beta0 + beta1 * STR + u
ca <- data.frame(STR = STR, TestScore = TestScore)

round(sapply(ca, function(v) c(media = mean(v), dp = sd(v),
                               min = min(v), max = max(v))), 2)
```

```
        STR TestScore
media 19.65    654.33
dp     1.91     19.09
min   13.58    591.29
max   25.28    699.86
```

Note o suporte de STR: de aproximadamente 13,6 a 25,3. O valor $STR = 0$, ao qual o intercepto se refere, está a mais de sete desvios-padrão do menor valor observado — a ilustração numérica exata do argumento sobre extrapolação feito na 1.

4.9.2 (a) A regressão

```
reg <- lm(TestScore ~ STR, data = ca)
round(coef(reg), 4)
```

```
(Intercept)         STR
   698.8918     -2.2682
```

```
round(c(R2 = summary(reg)$r.squared,
        SER = summary(reg)$sigma,
        n = n_d), 4)
```

```
      R2      SER        n
  0.0515  18.6166 420.0000
```

A reta estimada é $\widehat{TestScore} = 698{,}89 - 2{,}27 \cdot STR$, com $R^2 = 0{,}051$ e SER $= 18{,}62$. Comparando com o output do STATA no slide 34 (698,93, −2,2798, 0,0512, 18,581), a reconstrução é notavelmente próxima — o que era de se esperar, dado que os momentos foram calibrados para isso, mas serve para confirmar que a calibração está correta. A diferença entre −2,268 e −2,280 é ruído amostral de uma única amostra de 420 observações, e a seção (f) mostrará que ela é inteiramente compatível com o erro-padrão do estimador.

```
set.seed(99)
amostra_g <- ca[sample(nrow(ca), 60), ]
amostra_g$ajustado <- coef(reg)[1] + coef(reg)[2] * amostra_g$STR

ggplot(ca, aes(STR, TestScore)) +
  geom_point(colour = cz, alpha = 0.25, size = 1) +
  geom_segment(data = amostra_g,
               aes(x = STR, xend = STR, y = TestScore, yend = ajustado),
               colour = vm, linewidth = 0.3, alpha = 0.8) +
  geom_point(data = amostra_g, colour = az, size = 1.2) +
  geom_abline(intercept = coef(reg)[1], slope = coef(reg)[2],
              colour = az, linewidth = 0.9) +
  labs(x = "STR (alunos por professor)", y = "TestScore")
```

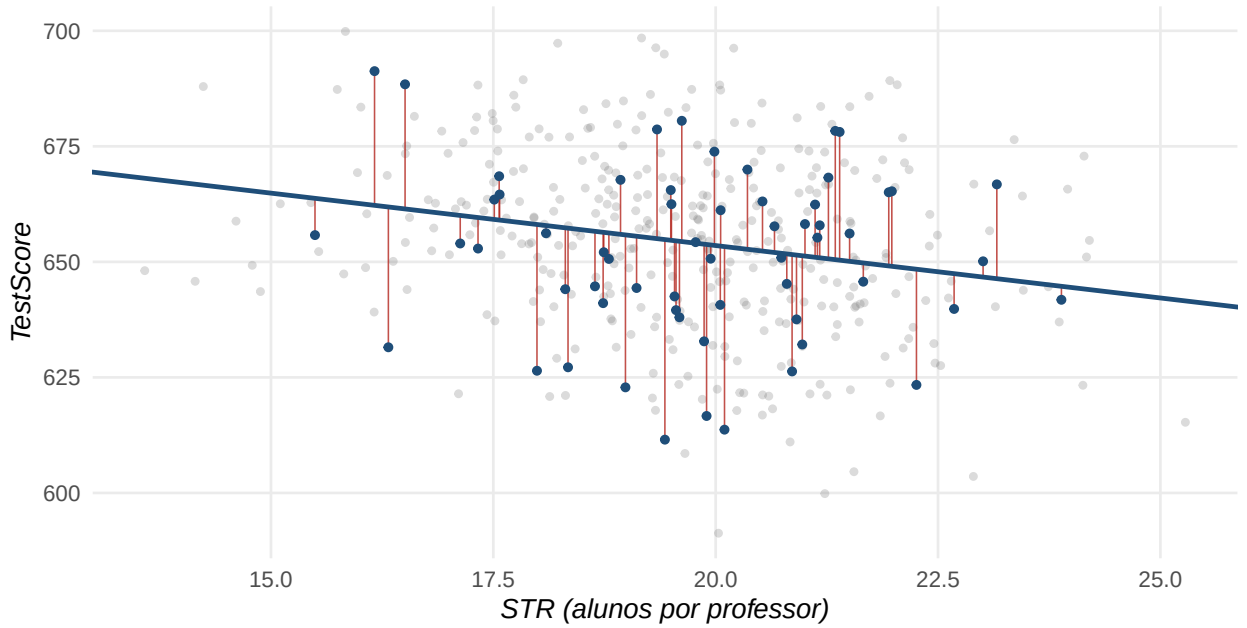

Figura 4: Nuvem de pontos e reta de MQO nos dados simulados. Os 420 distritos aparecem em cinza; uma subamostra aleatória de 60 está destacada em azul com seus resíduos $\hat{u}_i$ marcados em vermelho — a distância vertical entre a observação e a reta. O MQO escolhe a reta que minimiza a soma dos quadrados desses segmentos. Note que os segmentos são verticais, e não perpendiculares à reta: minimizamos o erro na previsão de Y, o que quebra a simetria entre as variáveis.

A figura torna visível o que o R^2 de 0,051 quer dizer: a reta tem inclinação negativa perceptível,

mas a nuvem é larga em torno dela. O tamanho da turma desloca a média condicional; está longe de determinar o desempenho de um distrito individual.

4.9.3 (b) MQO na mão contra `lm()`

Vamos verificar que a fórmula derivada algebricamente é o que o `lm()` calcula. Três caminhos: a fórmula com somatórios, a razão covariância-sobre-variância, e o `lm()`.

```
Xb <- mean(ca$STR);       Yb <- mean(ca$TestScore)
dX <- ca$STR - Xb;        dY <- ca$TestScore - Yb

b1_mao <- sum(dX * dY) / sum(dX^2)          # formula com somatorios
b0_mao <- Yb - b1_mao * Xb
b1_cov <- cov(ca$STR, ca$TestScore) / var(ca$STR)   # s_XY / s^2_X

round(rbind(formula_somatorios = c(b0 = b0_mao, b1 = b1_mao),
            cov_sobre_var      = c(b0 = NA,     b1 = b1_cov),
            lm                 = coef(reg)), 8)
```

```
                        b0        b1
formula_somatorios 698.8918 -2.268237
cov_sobre_var            NA -2.268237
lm                 698.8918 -2.268237
```

```
c(discrepancia_maxima = max(abs(c(b0_mao, b1_mao) - coef(reg))))
```

```
discrepancia_maxima
       1.136868e-13
```

As três linhas coincidem até a oitava casa decimal, e a discrepância máxima é da ordem de 10^{-13} — precisão de máquina em ponto flutuante, não diferença de método. As duas primeiras linhas conferem a álgebra dos Passos 1 a 4; a igualdade entre as duas primeiras confirma que $\widehat{\beta}_1 = s_{XY}/s_X^2$, com o $(n-1)$ se cancelando entre numerador e denominador exatamente como previsto.

4.9.4 (c) As condições de ortogonalidade

As duas identidades $\sum \hat{u}_i = 0$ e $\sum \hat{u}_i X_i = 0$ valem por construção. Vejamos.

```
uhat <- residuals(reg)
format(c(soma_u   = sum(uhat),
         soma_uX  = sum(uhat * ca$STR),
         media_u  = mean(uhat)), scientific = TRUE, digits = 3)
```

```
    soma_u    soma_uX    media_u
"3.77e-13" "1.35e-11" "1.04e-15"
```

As três quantidades são zero até o erro de arredondamento do ponto flutuante: 4×10^{-13} e 1×10^{-11} contra valores de Y da ordem de 650 e somas de quadrados da ordem de 10^5. Em termos relativos, são zeros exatos.

Vale insistir no que isso *não* significa. Que $\sum \hat{u}_i X_i = 0$ é uma propriedade do estimador, verdadeira mesmo se o modelo estiver completamente errado — mesmo se X for endógeno, se a relação for não linear, se os dados forem ruído. **Nenhum teste baseado nos resíduos pode detectar violação da Hipótese 1**, porque os resíduos são construídos para satisfazer a condição amostral correspondente. É a razão pela qual $\mathbb{E}(u \mid X) = 0$ não é testável.

4.9.5 (d) A decomposição TSS = ESS + SSR e as três formas do R^2

```
Yhat <- fitted(reg)
TSS  <- sum((ca$TestScore - Yb)^2)
ESS  <- sum((Yhat - Yb)^2)
SSR  <- sum(uhat^2)

round(c(TSS = TSS, ESS = ESS, SSR = SSR,
        ESS_mais_SSR = ESS + SSR,
        diferenca    = TSS - (ESS + SSR)), 6)
```

```
        TSS          ESS          SSR ESS_mais_SSR    diferenca
  152732.176     7863.176   144869.000   152732.176        0.000
```

A diferença é exatamente zero: a demonstração se confirma numericamente. Verifiquemos também que $\bar{\hat{Y}} = \bar{Y}$, o fato usado para justificar a definição de ESS:

```
c(media_Yhat = mean(Yhat), media_Y = Yb, diferenca = mean(Yhat) - Yb)
```

```
media_Yhat    media_Y  diferenca
  654.3315   654.3315     0.0000
```

Agora as três expressões do R^2 — as duas equivalentes da definição e a do item [a] do exercício do slide 33, $R^2 = r_{XY}^2$:

```
r_xy <- cor(ca$STR, ca$TestScore)
round(c(ESS_sobre_TSS    = ESS / TSS,
        um_menos_SSR_TSS = 1 - SSR / TSS,
        r2_da_correlacao = r_xy^2,
        R2_reportado     = summary(reg)$r.squared), 10)
```

```
   ESS_sobre_TSS um_menos_SSR_TSS r2_da_correlacao     R2_reportado
      0.05148343       0.05148343       0.05148343       0.05148343
```

As quatro coincidem até a décima casa. O item [a] está verificado. Vejamos os itens [b] e [c]:

```
reg_inv <- lm(STR ~ TestScore, data = ca)     # regressao de X em Y
round(c(R2_de_Y_em_X = summary(reg)$r.squared,
        R2_de_X_em_Y = summary(reg_inv)$r.squared), 10)          # item [b]
```

```
R2_de_Y_em_X R2_de_X_em_Y
  0.05148343   0.05148343
```

```
round(c(b1_do_lm     = unname(coef(reg)[2]),
        r_vezes_sY_sX = r_xy * sd(ca$TestScore) / sd(ca$STR)), 10)  # item [c]
```

```
     b1_do_lm r_vezes_sY_sX
    -2.268237     -2.268237
```

O item [b] se confirma: os dois R^2 são idênticos, embora as duas retas sejam completamente diferentes — a inclinação da regressão de X em Y é -0.0227, que **não** é o inverso de $-2{,}268$ (o inverso seria $-0{,}441$). A razão entre as duas inclinações é exatamente R^2: como $\hat{\beta}_1 = s_{XY}/s_X^2$ e $\tilde{\beta}_1 = s_{XY}/s_Y^2$, tem-se $\hat{\beta}_1 \cdot \tilde{\beta}_1 = s_{XY}^2/(s_X^2 s_Y^2) = r_{XY}^2 = R^2$. As duas retas só coincidem quando $R^2 = 1$. O ajuste é simétrico; os coeficientes não são. O item [c] também se confirma até a décima casa.

Por fim, o SER pela fórmula e a comparação com s_Y:

```
round(c(SER_pela_formula = sqrt(SSR / (n_d - 2)),
        SER_do_lm        = summary(reg)$sigma,
        sY               = sd(ca$TestScore),
        sY_vezes_raiz_1_menos_R2 = sd(ca$TestScore) *
          sqrt(1 - summary(reg)$r.squared)), 4)
```

```
        SER_pela_formula                 SER_do_lm                        sY
                 18.6166                   18.6166                   19.0923
sY_vezes_raiz_1_menos_R2
                 18.5943
```

SER $= 18{,}62$ contra $s_Y = 19{,}09$: condicionar em STR reduz a dispersão de 19,09 para 18,62 pontos, uma redução de menos de 3%. A quarta linha mostra a relação SER $\approx s_Y\sqrt{1-R^2}$, exata a menos da diferença entre os divisores $n-1$ e $n-2$. É a leitura absoluta do mesmo fato que o R^2 expressa em termos relativos.

4.9.6 (e) Erros-padrão robustos contra homocedásticos

Chegamos ao ponto em que o R reproduz o `robust` do STATA. A função `ep_robusto()`, definida no preâmbulo, implementa o estimador de White (HC1):

$$\widehat{\mathrm{Var}}_{\mathrm{HC1}}(\hat{\beta}) = \frac{n}{n-k}\,(\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{n}\hat{u}_i^{\,2}\,\mathbf{x}_i\mathbf{x}_i'\right)(\mathbf{X}'\mathbf{X})^{-1},$$

o chamado estimador "sanduíche" — duas fatias de $(\mathbf{X}'\mathbf{X})^{-1}$ com o recheio $\sum \hat{u}_i^2\mathbf{x}_i\mathbf{x}_i'$ no meio.

```
comparacao <- cbind(homocedastico = summary(reg)$coefficients[, 2],
                    robusto_HC1   = ep_robusto(reg))
round(comparacao, 4)
```

```
            homocedastico robusto_HC1
(Intercept)        9.3991      8.9701
STR                0.4762      0.4587
```

```
round(c(dif_pct_intercepto = 100 * (comparacao[1, 2] / comparacao[1, 1] - 1),
        dif_pct_STR        = 100 * (comparacao[2, 2] / comparacao[2, 1] - 1)), 2)
```

```
dif_pct_intercepto        dif_pct_STR
             -4.56              -3.67
```

Os dois erros-padrão **diferem** — o robusto é cerca de 4% menor para o coeficiente de STR. Por quê?

A fórmula homocedástica supõe $\mathrm{Var}(u_i \mid X_i) = \sigma_u^2$, constante para todo X_i, e sob essa hipótese a variância se reduz a $\mathrm{SER}^2/\sum(X_i - \bar{X})^2$. A fórmula robusta não supõe nada: ela estima $\mathrm{Var}[(X_i - \mu_X)u_i]$ diretamente, usando cada $\hat{u}_i^2$ como estimativa local da variância do erro naquele ponto. Se a homocedasticidade vale, as duas convergem para o mesmo valor e a diferença observada em amostras finitas é ruído — que é exatamente o caso aqui, já que **os dados foram simulados com erro homocedástico**. Os 4% de diferença são erro amostral na estimação do "recheio" do sanduíche.

A assimetria de risco, porém, é clara: se a homocedasticidade **falha** — e ela falha com frequência em dados econômicos, onde a dispersão do erro costuma crescer com o nível da variável —, o erro-padrão homocedástico é **inconsistente** e a inferência baseada nele é inválida, podendo errar em qualquer direção. O robusto é válido nos dois casos. Daí a regra prática, hoje universal em economia aplicada: **use sempre erros-padrão robustos**, e é por isso que o professor escreve `, robust` em todo comando `regress`.

A tabela completa, com estatísticas t e intervalos de confiança — cuja teoria é o assunto da Aula 4, mas cujo cálculo já podemos fazer:

```
round(tabela_reg(reg, robusto = TRUE), 4)
```

```
                coef     ep       t p   ic_inf   ic_sup
(Intercept) 698.8918 8.9701 77.9138 0 681.2597 716.5238
STR          -2.2682 0.4587 -4.9446 0  -3.1699  -1.3665
```

Compare com o output do STATA reproduzido acima: mesma estrutura, magnitudes compatíveis, e o intervalo de confiança para β_1 excluindo o zero com folga.

4.9.7 (f) A distribuição amostral de $\hat{\beta}_1$ por simulação

Esta é a figura central do capítulo. A teoria afirma que $\hat{\beta}_1$ é uma variável aleatória, não viesada, aproximadamente Normal em amostras grandes, com variância dada pela fórmula da caixa. Vamos **ver isso acontecer**.

O experimento é o que a definição de distribuição amostral literalmente descreve: repetir a amostragem muitas vezes, guardando a estimativa de cada amostra. Aqui podemos fazê-lo porque conhecemos o processo gerador — geramos 5.000 amostras de $n = 420$ do mesmo modelo, com $\beta_1 = -2{,}28$ verdadeiro, e coletamos os 5.000 valores de $\hat{\beta}_1$.

```r
set.seed(4231)
M <- 5000
b1_sim <- replicate(M, {
  Xs <- rnorm(n_d, mean = 19.64, sd = 1.89)
  Ys <- beta0 + beta1 * Xs + rnorm(n_d, mean = 0, sd = sigma_u)
  sum((Xs - mean(Xs)) * (Ys - mean(Ys))) / sum((Xs - mean(Xs))^2)
})

varX_pop    <- 1.89^2
var_teorica <- (varX_pop * sigma_u^2) / (n_d * varX_pop^2)   # = sigma_u^2 / (n Var(X))

round(c(media_das_estimativas = mean(b1_sim),
        beta1_verdadeiro      = beta1,
        vies_empirico         = mean(b1_sim) - beta1,
        dp_empirico           = sd(b1_sim),
        dp_teorico            = sqrt(var_teorica)), 5)
```

```
media_das_estimativas      beta1_verdadeiro         vies_empirico
              -2.28660               -2.28000               -0.00660
          dp_empirico            dp_teorico
              0.47789                0.47891
```

Os dois resultados centrais do capítulo aparecem nesses cinco números.

Não-viés. A média das 5.000 estimativas é $-2{,}2866$ contra o valor verdadeiro $-2{,}28$: um viés empírico de $-0{,}0066$, que é cerca de $1/70$ do desvio-padrão do estimador e perfeitamente compatível com o erro de Monte Carlo (o erro-padrão da média de 5.000 réplicas é $0{,}478/\sqrt{5000} \approx 0{,}0068$). $\mathbb{E}(\hat{\beta}_1) = \beta_1$, verificado.

A fórmula da variância. O desvio-padrão empírico das 5.000 estimativas é 0,4779; a fórmula teórica $\sqrt{\sigma_u^2/[n\mathrm{Var}(X)]}$ prevê 0,4789. Concordância na terceira casa decimal. Esse número também explica, retrospectivamente, a diferença entre a estimativa da seção (a), −2,268, e o valor verdadeiro −2,28: uma discrepância de 0,012, ou 0,025 desvio-padrão. Ruído amostral rotineiro.

```
d_sim <- data.frame(b1 = b1_sim)
ggplot(d_sim, aes(b1)) +
  geom_histogram(aes(y = after_stat(density)), bins = 50,
                 fill = az, colour = "white", alpha = 0.75, linewidth = 0.2) +
  stat_function(fun = dnorm,
                args = list(mean = beta1, sd = sqrt(var_teorica)),
                colour = vm, linewidth = 0.9) +
  geom_vline(xintercept = beta1, colour = cz, linetype = "dashed") +
  labs(x = "estimativas de beta1 nas 5.000 amostras", y = "densidade")
```

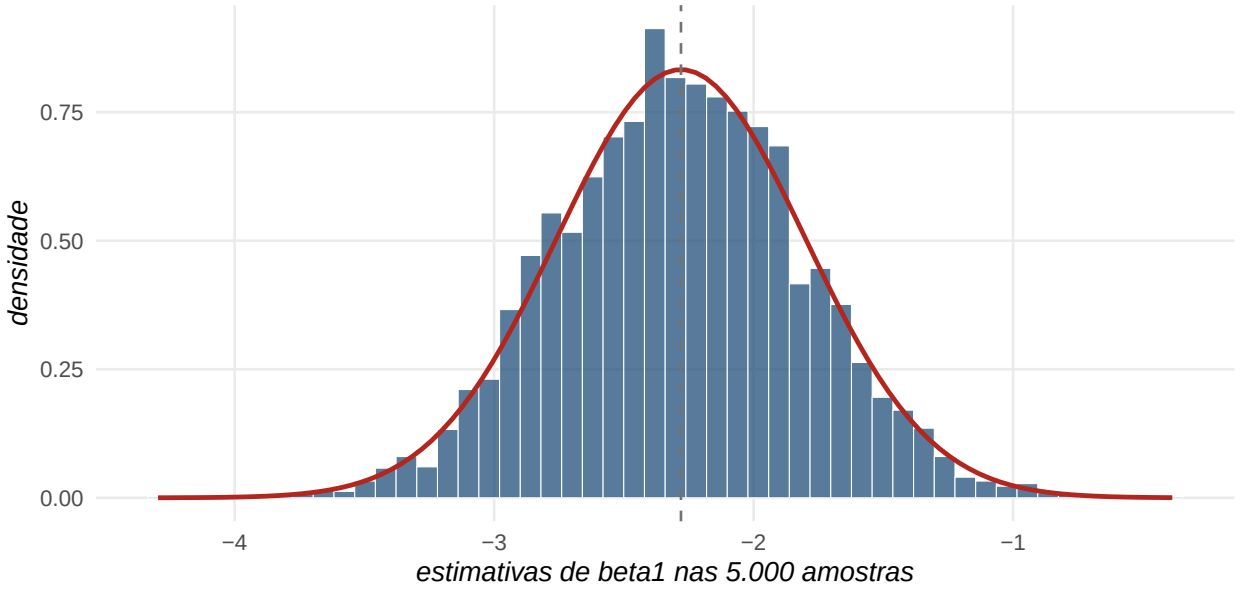

Figura 5: Distribuição amostral de $\hat{\beta}_1$: histograma das 5.000 estimativas obtidas por reamostragem do mesmo processo gerador ($n = 420$ em cada réplica). A curva vermelha é a densidade $\mathcal{N}(\beta_1, \sigma^2_{\hat{\beta}_1})$ prevista pela teoria, com variância dada pela fórmula de Stock e Watson; a linha cinza tracejada marca o verdadeiro $\beta_1 = -2{,}28$. O histograma é simétrico, tem forma de sino e se equilibra sobre o valor verdadeiro — não-viés e normalidade assintótica, vistos por simulação.

A figura é a tradução gráfica das duas propriedades. O histograma tem **forma de sino** — apesar de $\hat{\beta}_1$ ser uma razão de somas, objeto sem nenhuma razão óbvia para ser Normal — e a curva teórica se sobrepõe a ele quase perfeitamente. É o Teorema Central do Limite em ação, operando sobre o numerador $\frac{1}{n}\sum(X_i - \bar{X})u_i$. E o histograma **se equilibra sobre a linha tracejada**, que marca o β_1 verdadeiro: é o não-viés.

Vale notar o que a figura torna palpável. Em uma aplicação real observamos **um** ponto desse histograma — uma amostra, uma estimativa — e nunca saberemos qual. O erro-padrão é nossa

estimativa da largura dessa distribuição a partir daquele único ponto, e é isso que permite dizer algo sobre a distância provável entre a estimativa e o alvo. Toda a inferência da Aula 4 é construída sobre essa ideia.

Finalmente, os três determinantes da precisão. A tabela varia n e o desvio-padrão de X, mantendo σ_u fixo, e reporta $\mathrm{SD}(\hat{\beta}_1) = \sigma_u/[\sqrt{n} \cdot \mathrm{sd}(X)]$:

```
grade <- expand.grid(n = c(100, 420, 2000), sd_X = c(0.9, 1.89, 4))
grade$dp_beta1 <- with(grade, sigma_u / (sqrt(n) * sd_X))
round(grade[order(grade$sd_X, grade$n), ], 4)
```

```
     n sd_X dp_beta1
1  100 0.90   2.0611
2  420 0.90   1.0057
3 2000 0.90   0.4609
4  100 1.89   0.9815
5  420 1.89   0.4789
6 2000 1.89   0.2195
7  100 4.00   0.4638
8  420 4.00   0.2263
9 2000 4.00   0.1037
```

Leia a tabela em duas direções. **Descendo** dentro de cada bloco de `sd_X`: multiplicar n por 20 (de 100 para 2000) divide o erro-padrão por $\sqrt{20} \approx 4{,}5$. **Comparando blocos** com o mesmo n: quadruplicar a dispersão do regressor (de 0,9 para 4) divide o erro-padrão por 4 — efeito **proporcional**, mais barato que aumentar a amostra. É o argumento econométrico a favor de desenhos que gerem **variação no regressor**, e a razão pela qual um experimento com forte contraste entre tratamento e controle é mais informativo que um com contraste tênue, ainda que ambos tenham a mesma amostra.

4.9.8 (g) O que uma única observação extrema pode fazer

Para fechar, uma ilustração da Hipótese 3. Geramos 40 observações bem comportadas de uma relação com inclinação 1,2 e acrescentamos **uma** observação extrema.

```
set.seed(7)
n_o <- 40
Xo  <- rnorm(n_o, mean = 10, sd = 2)
Yo  <- 5 + 1.2 * Xo + rnorm(n_o, mean = 0, sd = 1.2)
d_lim <- data.frame(X = c(Xo, 20), Y = c(Yo, 2))    # a ultima linha e o valor extremo

m_sem <- lm(Y ~ X, data = d_lim[1:n_o, ])
m_com <- lm(Y ~ X, data = d_lim)
round(rbind(sem_o_extremo = coef(m_sem),
            com_o_extremo = coef(m_com)), 4)
```

```
              (Intercept)      X
sem_o_extremo      3.4695 1.3591
com_o_extremo     13.8089 0.3344
```

O coeficiente angular cai de 1,36 para 0,33 — uma redução de 75% — por causa de **uma** observação em 41. O sinal foi preservado por sorte; com um valor extremo um pouco mais distante, teria invertido. É a ilustração exata do argumento do professor: o MQO minimiza uma soma de **quadrados**, e uma observação distante contribui com o quadrado de sua distância, dominando o critério. Nenhum teste estatístico substitui, aqui, o hábito de olhar para os dados.

4.10 Exercícios

1. **Álgebra do estimador.** Considere a regressão $Y_i = \beta_0 + \beta_1 X_i + u_i$ estimada por MQO. (a) Mostre que a reta de MQO passa necessariamente por $(\bar{X}, \bar{Y})$. (b) Mostre que $\sum_i \hat{u}_i \hat{Y}_i = 0$, isto é, os resíduos são ortogonais também aos valores ajustados. (c) Suponha que se estime a regressão **sem intercepto**, $\min_{b_1} \sum (Y_i - b_1 X_i)^2$. Obtenha a CPO e mostre que $\tilde{\beta}_1 = \sum X_i Y_i / \sum X_i^2$. (d) Nesse caso, ainda vale $\sum \hat{u}_i = 0$? O que isso implica para a validade da decomposição TSS = ESS + SSR e para a interpretação do R^2?

2. **Mudança de unidades.** A regressão $\widehat{TestScore} = 698{,}9 - 2{,}28\, STR$ foi estimada com $TestScore$ em pontos. (a) Se o teste for reescalado para uma nota de 0 a 10 (dividindo por 80), o que acontece com $\hat{\beta}_0$, $\hat{\beta}_1$, R^2 e SER? (b) E se em vez disso somarmos 50 pontos a todas as notas? (c) E se STR for medido em "professores por 100 alunos" em lugar de "alunos por professor" — a transformação é linear? (d) Enuncie a regra geral: sob $Y^* = a + bY$ e $X^* = c + dX$, como se transformam os quatro objetos? Qual deles é **invariante** e por quê?

3. **A decomposição, à mão.** Uma amostra de $n = 5$ observações produziu: $\sum (Y_i - \bar{Y})^2 = 200$, $\sum \hat{u}_i^2 = 50$. (a) Calcule ESS, R^2 e SER.

   (b) Qual é a correlação amostral entre X e Y, em valor absoluto? (c) Sabendo que $s_X = 2$ e que a relação é positiva, obtenha $\hat{\beta}_1$. (d) Por que a informação dada não permite recuperar $\hat{\beta}_0$?

4. **A hipótese de média condicional zero.** Em cada caso abaixo, discuta se $\mathbb{E}(u_i \mid X_i) = 0$ é plausível, identificando o fator omitido relevante e o **sinal** provável do viés: (a) regressão do salário sobre anos de escolaridade; (b) regressão da taxa de criminalidade municipal sobre o número de policiais por habitante; (c) regressão do desempenho de um fundo sobre a taxa de administração que ele cobra; (d) regressão do peso ao nascer sobre o número de consultas pré-natais. Em qual deles você imaginaria um experimento aleatorizado viável, e o que ele aleatorizaria?

5. **Correlação zero não basta.** Seja $X \sim \mathcal{N}(0, 1)$ e $u = X^2 - 1$. (a) Verifique que $\mathbb{E}(u) = 0$ e $\mathrm{Cov}(u, X) = 0$. (b) Verifique que $\mathbb{E}(u \mid X) \neq 0$. (c) Gere esses dados em R com $n = 10.000$, rode $Y = 1 + 2X + u$ contra X e verifique que $\hat{\beta}_1$ é próximo de 2 — isto é, o estimador **não** é viesado neste caso particular. (d) Explique por quê: qual condição, mais fraca que $\mathbb{E}(u \mid X) = 0$, é suficiente para o não-viés assintótico de $\hat{\beta}_1$? (e) Construa agora um exemplo com u correlacionado com X e mostre, por simulação, que $\hat{\beta}_1$ passa a ser viesado.

6. **O beta do CAPM.** Considere a regressão $R_{it} - R_{ft} = \alpha_i + \beta_i (R_{mt} - R_{ft}) + u_{it}$. (a) Mostre que, se o CAPM vale exatamente, $\alpha_i = 0$. (b) Interprete o R^2 dessa regressão em termos da decomposição do risco entre sistemático e idiossincrático. (c) Um analista afirma que, como o R^2 da regressão de uma ação foi de apenas 0,25, "o beta estimado não presta". Avalie a afirmação, distinguindo entre o valor do R^2 e a precisão de $\widehat{\beta}_i$. (d) Que propriedade da série de retornos torna a Hipótese 2 delicada aqui, e o que se costuma fazer a respeito?

7. **O R^2 e a causalidade.** (a) Construa um exemplo numérico (pode ser em R) de duas variáveis com $R^2 > 0{,}95$ e nenhuma relação causal entre elas. (b) Construa um exemplo de uma relação causal genuína com $R^2 < 0{,}05$. (c) Por que acrescentar um regressor **nunca** reduz o R^2, mesmo que o regressor seja ruído puro? (d) Que estatística tenta corrigir esse problema, e por que ela também não resolve a questão causal?

8. **Simulação.** Adapte o código da seção (f) do laboratório. (a) Reduza n para 20 e refaça o histograma: a aproximação Normal ainda é boa? (b) Mantenha $n = 20$ mas gere os erros de uma distribuição t de Student com 3 graus de liberdade (que tem quarto momento **infinito**, violando a Hipótese 3): o que acontece com as caudas da distribuição amostral? (c) Volte à Normal e faça $\mathrm{sd}(X)$ variar em $\{0{,}5,\ 1,\ 2,\ 4\}$, verificando que $\mathrm{SD}(\widehat{\beta}_1)$ cai proporcionalmente a $1/\mathrm{sd}(X)$.

   (d) Introduza heterocedasticidade fazendo $\mathrm{Var}(u_i \mid X_i) = (0{,}5\,X_i)^2$ e compare, ao longo das réplicas, o desvio-padrão empírico de $\widehat{\beta}_1$ com a média dos erros-padrão homocedásticos e a média dos robustos. Qual dos dois acerta?

5 Inferência na Regressão com um Regressor

O 1 terminou com uma dívida. Diante da saída do STATA que o professor projeta desde a Aula 3, sabíamos ler quase tudo — `Coef.` são $\hat{\beta}_0$ e $\hat{\beta}_1$, `R-squared` é ESS/TSS, `Root MSE` é o SER —, mas paramos na coluna `Std. Err.` e em tudo o que dela decorre: as estatísticas t, os p-valores, os intervalos de confiança. Este capítulo paga essa dívida.

A razão de a coluna existir já está estabelecida: $\hat{\beta}_1$ é uma **variável aleatória**, porque é calculado a partir de uma amostra aleatória. Se tivéssemos sorteado outros 420 distritos, teríamos obtido outro número. A pergunta que organiza o capítulo é o que fazer com essa incerteza:

> Dada uma estimativa $\hat{\beta}_1 = -2{,}28$ obtida em *uma* amostra, o que podemos afirmar sobre o parâmetro populacional β_1, que não observamos?

Há duas maneiras de responder, e elas são duas faces do mesmo objeto. A primeira é **testar hipóteses**: confrontar um valor conjecturado para β_1 com a evidência amostral e decidir se ela o contradiz. A segunda é construir um **intervalo de confiança**: exibir o conjunto de valores de β_1 que a amostra *não* rejeita. Ao final da Seção 5 veremos que as duas coisas são literalmente equivalentes.

O capítulo corresponde à Aula 4 da disciplina e fecha o tratamento do Capítulo 4 de Stock e Watson (2020), avançando sobre o Capítulo 5. Sua estrutura é a do professor: uma revisão de testes de hipótese em termos gerais; a aplicação aos coeficientes da regressão; os intervalos de confiança; o caso em que o regressor é **binário** — que revela a regressão como uma comparação de médias; a **heterocedasticidade** e suas consequências para a inferência; e, por fim, o **Teorema de Gauss-Markov**, apresentado com o ceticismo que o professor faz questão de registrar já no título do slide, escrito no passado.

5.1 Revisão de testes de hipótese

Antes de tratar dos coeficientes da regressão, o professor recupera a lógica do teste de hipóteses em termos gerais, com um parâmetro populacional genérico θ. Vale segui-lo: a estrutura é a mesma que reaparecerá aplicada a β_1, e é mais fácil vê-la primeiro sem a bagagem da regressão.

A informação da amostra pode ser usada para avaliar a validade de uma **conjectura** sobre um parâmetro populacional. Formula-se uma hipótese sobre θ, que se considera válida **a menos que evidência contrária seja produzida** — essa é a **hipótese nula**, H_0. Se ela não é válida, uma **hipótese alternativa** H_1 deve ser verdadeira.

A hipótese nula pode ser **simples**, quando fixa um único valor,

$$H_0 : \theta = \theta_0,$$

ou **composta**, quando especifica uma faixa de valores:

$$H_0 : \theta \geq 20 \qquad \text{contra} \qquad H_1 : \theta < 20.$$

Especificadas as hipóteses e coletada a informação amostral, é preciso decidir. Há duas possibilidades — **aceitar** H_0 ou **rejeitá-la** em favor de H_1 — e a decisão exige uma **regra** baseada na evidência amostral.

> **Sobre o verbo "aceitar"**
>
> O professor usa "aceitar a hipótese nula", e mantenho o vocabulário dele. Vale registrar, porém, a razão pela qual boa parte da literatura prefere **"não rejeitar"**: um teste que não rejeita H_0 não produziu evidência de que H_0 seja verdadeira; produziu apenas evidência insuficiente para descartá-la. A assimetria é a mesma do júri, exemplo que o próprio professor usa adiante — absolver não é declarar inocente, é não ter provado a culpa. Essa distinção volta com força na discussão do **poder do teste**, logo abaixo: quando o poder é baixo, "não rejeitar" é quase automático e, portanto, quase não informa.

5.1.1 Os dois tipos de erro

Como a decisão se baseia em uma amostra, ela pode errar — e há duas maneiras de errar, que não são simétricas. O professor as organiza na tabela de probabilidades condicionais:

	H_0 é verdadeira	H_0 é falsa
Aceitar H_0	Decisão correta $(1-\alpha)$	**Erro Tipo II** (β)
Rejeitar H_0	**Erro Tipo I** (α)	Decisão correta $(1-\beta)$

O **erro do Tipo I** é rejeitar uma hipótese nula verdadeira. Sua probabilidade é o **nível de significância** do teste:

$$\boxed{\alpha = \Pr(\text{Rejeitar } H_0 \mid H_0 \text{ é verdadeira})}$$

O **erro do Tipo II** é aceitar uma hipótese nula falsa:

$$\boxed{\beta = \Pr(\text{Aceitar } H_0 \mid H_0 \text{ é falsa})}$$

> **Notação — dois β diferentes**
>
> O símbolo β designa, neste capítulo, duas coisas distintas: a probabilidade do erro do Tipo II, na notação clássica de testes, e os **coeficientes da regressão** β_0, β_1. É uma colisão de notação herdada da literatura, e o contexto sempre resolve — β sozinho, nesta seção, é probabilidade de erro; com índice, é coeficiente.

O exemplo do teste de gravidez. Testes de gravidez têm como hipótese nula "não há gravidez" (mais precisamente, que o nível do hormônio hCG está no limite basal normal). Então o erro do Tipo I,

$$\Pr(\text{Rejeita ausência de gravidez} \mid \text{não há gravidez}),$$

é o **falso positivo**, e o erro do Tipo II é o **falso negativo**. Vale notar como a identificação de qual erro é qual depende inteiramente de como se escolheu a nula: trocada a nula, trocam-se os rótulos.

O exemplo do júri. No sistema judicial, a hipótese nula é a **presunção de inocência**, e o critério de rejeição é que as evidências estejam "além de qualquer dúvida razoável" (*beyond reasonable doubt*). Nesse caso,

$$\alpha = \Pr(\text{Condenar um inocente} \mid \text{Inocente}), \qquad \beta = \Pr(\text{Aceitar inocência} \mid \text{Culpado}).$$

Os erros são **assimétricos em suas implicações**: é melhor absolver um culpado do que condenar um inocente. Por isso impõe-se um α pequeno — 1% ou menos, no espírito da dúvida razoável.

5.1.2 O trade-off entre os dois erros

Aqui está o ponto que o professor faz questão de desenvolver. Dado o tamanho da amostra, uma vez escolhido α, o valor de β **está dado**. Não é possível reduzir os dois ao mesmo tempo.

A intuição é direta e vale reproduzi-la. Imagine que se deseje $\alpha = 0$, isto é, nunca condenar um inocente. A única forma de garantir isso é **nunca rejeitar** H_0 — absolver sempre, qualquer que seja a evidência. Mas então $\beta = 100\%$: todo culpado é absolvido, e o teste é inútil.

$$\boxed{\text{Dado } n, \text{ reduzir } \alpha \implies \text{aumentar } \beta}$$

A única forma de reduzir **ambos** os erros é **aumentar o tamanho da amostra**. Mais informação desloca a fronteira inteira; sem mais informação, só se anda sobre ela.

5.1.3 O poder do teste

O complemento do erro do Tipo II tem nome próprio. O **poder** (ou potência) do teste é a probabilidade de rejeitar H_0 quando ela é de fato falsa:

$$\boxed{\text{Poder} = 1 - \beta = \Pr(\text{Rejeitar } H_0 \mid H_0 \text{ é falsa})}$$

O professor acrescenta um alerta que é de economia política da pesquisa, e não de estatística: testes com **baixo poder** aceitam a hipótese nula com frequência alta, o que **pode ser explorado pelo pesquisador interessado em aceitar** H_0. Quem quer concluir que um efeito não existe tem, num teste de baixo poder, um aliado silencioso — basta usar uma amostra pequena e apresentar a não rejeição como se fosse evidência de nulidade. É a razão pela qual "não encontramos efeito significativo" é uma frase que só se pode avaliar sabendo qual efeito o desenho teria sido capaz de detectar.

5.1.4 Um exemplo completo: teste sobre a média populacional

Para fixar o procedimento, o professor faz o teste mais simples possível — sobre a média de uma população — antes de passar aos coeficientes da regressão.

Sabemos das Notas de Estatística que o estimador da média populacional é a média amostral $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, e que, assumindo normalidade na população,

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$

Considere o teste **unicaudal**

$$H_0 : \mu = 5 \qquad \text{contra} \qquad H_1 : \mu > 5,$$

com $\sigma = 0{,}1$, $n = 16$, $\alpha = 5\%$ e média amostral observada $\bar{x} = 5{,}038$. A regra de decisão é rejeitar H_0 se

$$\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha, \qquad \text{onde} \qquad \Pr(Z > z_\alpha) = 0{,}05 \implies z_\alpha = 1{,}645.$$

O número $z_\alpha = 1{,}645$ é o **valor crítico**. Calculando a estatística de teste:

$$\frac{5{,}038 - 5}{0{,}1/\sqrt{16}} = \frac{0{,}038}{0{,}025} = 1{,}52.$$

Como $1{,}52 < 1{,}645$, a estatística está à esquerda do valor crítico, dentro da região de aceitação. **Conclusão:** com base na evidência amostral, não rejeitamos a possibilidade de a média populacional ser igual a 5.

Note que a estrutura tem quatro peças, e todas reaparecerão na regressão: (i) uma hipótese nula que fixa um valor; (ii) um estimador com distribuição amostral conhecida; (iii) uma **estatística de teste** que mede a distância entre a estimativa e o valor hipotético, em unidades de erro-padrão; e (iv) um valor crítico que define quão longe é longe demais.

5.2 Testando hipóteses sobre os coeficientes da regressão

Com a estrutura montada, a aplicação à regressão é quase imediata. Retomamos o exemplo da Aula 3,

$$\widehat{TestScore} = 698{,}9 - 2{,}28 \cdot STR,$$

e a pergunta empírica: o coeficiente de STR é **estatisticamente significante**? Isto é, queremos testar

$$H_0 : \beta_1 = 0 \qquad \text{contra} \qquad H_1 : \beta_1 \neq 0,$$

onde a alternativa é **bicaudal** — nos interessa detectar desvios de zero em qualquer direção.

> **!** Por que $\beta_1 = 0$ é a nula interessante
>
> A nula $\beta_1 = 0$ é a hipótese de que X **não tem efeito sobre** Y — de que a PRF é horizontal e conhecer o tamanho da turma não altera em nada a previsão do desempenho. É a hipótese que o pesquisador em geral quer derrubar, e é por isso que o software a testa por padrão, imprimindo a coluna de t e de p-valores para essa nula específica.
> Nada impede, porém, testar $H_0 : \beta_1 = \beta_{1,0}$ para qualquer outro valor conjecturado — e há contextos em que a nula relevante não é zero. No CAPM discutido no capítulo anterior, a hipótese teórica de interesse sobre o beta de uma ação é $H_0 : \beta_i = 1$ (a ação acompanha o mercado), e sobre o alfa é $H_0 : \alpha_i = 0$. A fórmula geral abaixo cobre todos esses casos; o software é que privilegia um deles.

5.2.1 A estatística t

Construímos a **estatística-t** exatamente como no teste sobre a média — a distância entre a estimativa e o valor hipotético, medida em erros-padrão:

$$t = \frac{\text{estimador} - \text{valor hipotético}}{\text{erro-padrão do estimador}} = \frac{\hat{\beta}_1 - \beta_{1,0}}{\text{SE}(\hat{\beta}_1)}$$

O denominador é o **erro-padrão** de $\hat{\beta}_1$: a raiz da estimativa da variância $\sigma^2_{\hat{\beta}_1}$ cuja fórmula fechamos no capítulo anterior. É a coluna `Std. Err.` do STATA, e agora sabemos para que ela serve.

A justificativa para comparar t com valores críticos da Normal está inteiramente contida no resultado de normalidade assintótica obtido no 1: sob as três hipóteses de MQO, $\hat{\beta}_1 \overset{a}{\sim} \mathcal{N}(\beta_1, \sigma^2_{\hat{\beta}_1})$ em amostras grandes. Padronizando sob a nula, $t \overset{a}{\sim} \mathcal{N}(0,1)$.

5.2.2 Aplicando ao exemplo

Retomamos a saída do STATA:

```
regress testscr str, robust

Regression with robust standard errors                Number of obs =     420
                                                      F(  1,   418) =   19.26
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.0512
                                                      Root MSE      =  18.581
------------------------------------------------------------------------------
             |               Robust
testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------+---------------------------------------------------------------------
```

```
       str |  -2.279808   .5194892    -4.39   0.000    -3.300945   -1.258671
     _cons |    698.933   10.36436    67.44   0.000     678.5602    719.3057
------------------------------------------------------------------------------
```

A estatística t do coeficiente de STR é

$$t = \frac{-2{,}28 - 0}{0{,}5195} = -4{,}39,$$

que é exatamente a coluna `t` da saída. O valor crítico para um teste bicaudal a 5% vem de

$$\Pr(-1{,}96 < Z < 1{,}96) = 0{,}95,$$

e como $-4{,}39$ está **fora** do intervalo $(-1{,}96;\ 1{,}96)$, **rejeitamos** H_0. O efeito estimado do tamanho da turma sobre o desempenho é estatisticamente significante ao nível de 5%.

> **i** Significância estatística não é significância econômica
>
> Rejeitar $H_0 : \beta_1 = 0$ diz que o efeito **não é zero**; não diz que ele é **grande**. São perguntas diferentes, e a segunda é a que interessa para política pública.
> Aqui: reduzir STR em duas unidades eleva o $TestScore$ em 4,56 pontos. Como o desvio-padrão dos test scores na amostra é de cerca de 19 pontos, isso é aproximadamente **um quarto de desvio-padrão** — modesto, e comprado ao custo de contratar professores. Com n suficientemente grande, efeitos economicamente irrelevantes tornam-se estatisticamente significantes; a estatística t cresce com $\sqrt{n}$ enquanto a magnitude do efeito não muda. Reportar apenas estrelas de significância, sem discutir magnitudes, é um vício comum e um erro de leitura.

5.2.3 O valor-p

Na tabela de regressão aparece a coluna $P > |t|$. O **valor-p** é o menor nível de significância α ao qual se rejeita H_0.

Voltemos ao teste sobre a média para ver de onde isso vem. Lá, a estatística era 1,52, o teste era unicaudal à direita, e escolhemos $\alpha = 5\%$; como 1,52 ficou abaixo do valor crítico $z_\alpha = 1{,}645$, não rejeitamos H_0. Para *rejeitar*, teríamos precisado de um α grande o bastante para que o valor crítico caísse abaixo de 1,52 — isto é, um α de pelo menos

$$\Pr(Z > 1{,}52) = 0{,}0643.$$

Esse número, 6,43%, **é** o valor-p. Dele decorre a regra de decisão:

rejeita-se H_0 sempre que $\alpha \geq$ valor-p;	aceita-se H_0 sempre que $\alpha <$ valor-p

No caso bicaudal, que é o da tabela de regressão, é preciso somar as duas caudas:

$$\text{valor-p} = \Pr(Z < -4{,}39) + \Pr(Z > 4{,}39),$$

e, por simetria da Normal,

$$\boxed{\text{valor-p} = 2 \cdot F_Z(-|t|) = 2 \cdot F_Z(-4{,}39)}$$

que é um número da ordem de 10^{-5} — impresso como `0.000` na saída do STATA. A vantagem do valor-p sobre a comparação com valor crítico é que ele **não exige fixar α de antemão**: reporta-se o valor-p e cada leitor aplica o seu próprio limiar.

5.3 Intervalos de confiança

Qualquer estimativa de β_1 contém, necessariamente, incerteza amostral — não podemos determinar o verdadeiro valor de β_1 a partir de uma amostra. O que podemos fazer é construir um **intervalo de confiança**, a partir da estimativa de MQO e do seu erro-padrão.

5.3.1 A definição geral

O professor começa pela definição geral de estimador intervalar, que vale a pena enunciar com cuidado porque a interpretação depende dela.

> **!** Definição — estimador intervalar
>
> Um **estimador intervalar** de um parâmetro populacional é uma regra para determinar, com base na informação amostral, um intervalo no qual é provável que o parâmetro populacional se situe.
> Formalmente: seja θ o parâmetro a ser estimado. Se, a partir da informação amostral, podemos encontrar duas **variáveis aleatórias** $A < B$ tais que
>
> $$\Pr(A < \theta < B) = 0{,}90,$$
>
> então o intervalo entre A e B é um estimador intervalar de θ com **90% de confiança**.

O ponto crucial está em quem é aleatório: A e B são variáveis aleatórias, funções da amostra; θ é um número fixo e desconhecido. A probabilidade é sobre o **intervalo**, não sobre o parâmetro.

5.3.2 Construção: o caso da média

Vamos construir o intervalo para a média populacional μ, onde toda a álgebra fica visível. Sabemos que $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ e, portanto, que

$$\Pr\big(-z_{\alpha/2} < Z < z_{\alpha/2}\big) = 1-\alpha, \qquad \text{por exemplo} \qquad \Pr(-1{,}645 < Z < 1{,}645) = 0{,}90.$$

Substituindo Z e isolando μ no meio da desigualdade:

$$\Pr\left(-1{,}645 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1{,}645\right) = 0{,}90$$

$$\Pr\left(\bar{X} - 1{,}645 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1{,}645 \cdot \frac{\sigma}{\sqrt{n}}\right) = 0{,}90,$$

e, de forma mais geral,

$$\boxed{\Pr\left(\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha}$$

Dizemos que a probabilidade de o **intervalo aleatório** conter a média populacional é $1-\alpha$. Uma vez que $\bar{X}$ é "realizado", isto é, se torna uma estimativa $\bar{x}$, obtemos uma **estimativa intervalar**:

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

Numericamente, com $\sigma = 0{,}1$, $n = 16$, $\bar{x} = 5{,}038$ e $1-\alpha = 0{,}95$:

$$5{,}038 \pm 1{,}96 \cdot \frac{0{,}1}{\sqrt{16}} \;=\; 5{,}038 \pm 0{,}049 \;\longrightarrow\; (4{,}989;\ 5{,}087).$$

5.3.3 Como interpretar — e como não interpretar

O professor dedica três slides à interpretação, e o cuidado se justifica porque o erro é extremamente comum.

É tentador ler (4,989; 5,087) como "o parâmetro populacional está contido no intervalo com 95% de probabilidade". Essa leitura de fato aparece com frequência, mas é preciso entender em que sentido ela se aplica — porque **a probabilidade de um número fixo estar contido em um intervalo fixo é zero ou 100%**. Uma vez realizado o intervalo, não há mais nada de aleatório: ou μ está dentro, ou não está. Nós é que não sabemos qual.

> **!** A interpretação formal
>
> Em **amostras repetidas**, usando a fórmula $\bar{x} \pm 1{,}96 \cdot \sigma/\sqrt{n}$ para calcular o intervalo, **95% dos intervalos conterão o parâmetro populacional**.
> A aleatoriedade está no **procedimento**, não no parâmetro. O que tem cobertura de 95% é a regra de construção, aplicada repetidamente a amostras diferentes — cada uma produzindo um intervalo diferente, dos quais 95% acertam o alvo fixo.

O laboratório deste capítulo constrói exatamente esse experimento: simula muitas amostras, calcula o intervalo em cada uma e conta a fração que contém o verdadeiro β_1.

5.3.4 O intervalo de confiança para β_1

Seguindo a mesma lógica, e sabendo que em amostras grandes

$$\hat{\beta}_1 \overset{a}{\sim} \mathcal{N}\left(\beta_1,\, \sigma^2_{\hat{\beta}_1}\right),$$

o intervalo de confiança de 95% para β_1 é

$$\boxed{\hat{\beta}_1 \pm 1{,}96 \cdot \mathrm{SE}(\hat{\beta}_1)}$$

No exemplo:

$$-2{,}28 \pm 1{,}96 \cdot 0{,}5195 \;\longrightarrow\; (-3{,}2982;\; -1{,}2618),$$

que é precisamente a coluna `[95% Conf. Interval]` da saída do STATA.

A interpretação é a já discutida: em amostras repetidas, 95% dos intervalos construídos por essa fórmula conterão o parâmetro populacional. E há uma leitura adicional, que o professor destaca: **o intervalo não contém o zero**. Logo, o valor $\beta_1 = 0$ pode ser rejeitado ao nível de significância de 5%.

> **!** A dualidade entre teste e intervalo
>
> Isso não é coincidência. O intervalo de confiança de 95% é **exatamente o conjunto de valores de $\beta_{1,0}$ que o teste bicaudal a 5% não rejeita**:
>
> $$\left|\frac{\hat{\beta}_1 - \beta_{1,0}}{\mathrm{SE}(\hat{\beta}_1)}\right| \leq 1{,}96 \quad \Longleftrightarrow \quad \beta_{1,0} \in \left[\hat{\beta}_1 - 1{,}96\,\mathrm{SE}(\hat{\beta}_1),\; \hat{\beta}_1 + 1{,}96\,\mathrm{SE}(\hat{\beta}_1)\right].$$
>
> As duas expressões são a mesma desigualdade, com $\beta_{1,0}$ isolado de um lado ou do outro. Daí a regra prática: **o intervalo de 95% contém zero se e somente se o teste de significância a 5% não rejeita**. O intervalo é mais informativo, porque exibe de uma vez todos os valores compatíveis com os dados, em vez de responder sim ou não sobre um único deles.

5.3.5 Intervalos para efeitos preditos

Raramente interessa uma mudança unitária em X. Em geral queremos o efeito de uma mudança de tamanho Δx — reduzir a turma em dois alunos, aumentar a escolaridade em quatro anos. A mudança esperada em Y é $\beta_1 \Delta x$, e como Δx é uma **constante**, o intervalo se obtém simplesmente multiplicando:

$$\boxed{\left(\hat{\beta}_1 \pm 1{,}96 \cdot \mathrm{SE}(\hat{\beta}_1)\right) \Delta x}$$

Exemplo. O IC para β_1 é $(-3{,}2982;\, -1{,}2618)$. Considere uma redução da razão aluno-professor de duas unidades, $\Delta STR = -2$. O efeito pontual estimado é

$$\Delta \widehat{TestScore} = -2{,}28 \cdot (-2) = 4{,}56,$$

e o intervalo de confiança é

$$(-3{,}2982;\ -1{,}2618) \cdot (-2) \ \longrightarrow\ (2{,}5236;\ 6{,}5964).$$

Note a inversão da ordem dos extremos: multiplicar por um número **negativo** troca o sinal das desigualdades, de modo que o limite inferior do IC de β_1 vira o limite superior do IC do efeito. Com 95% de confiança, portanto, reduzir a turma em dois alunos eleva o desempenho entre 2,5 e 6,6 pontos.

5.4 Regressão quando X é binário

Até aqui o regressor foi uma variável contínua. Mas a análise de regressão produz resultados interessantes — e conceitualmente esclarecedores — quando o regressor é **binário**, assumindo apenas os valores 0 ou 1. É o caso de uma *dummy* como `female`, que vale 1 se o indivíduo é mulher e 0 caso contrário.

Considere o modelo

$$Y_i = \beta_0 + \beta_1 D_i + u_i, \qquad i = 1, \ldots, n,$$

com D_i binária. No exemplo do capítulo:

$$D_i = \begin{cases} 1, & \text{se } STR \text{ no distrito } i < 20 \\ 0, & \text{se } STR \text{ no distrito } i \geq 20 \end{cases}$$

Como interpretar β_1? Note, antes de tudo, que a leitura usual — "a variação em Y associada a uma variação unitária em X" — perde o sentido literal aqui: não existe "metade de uma unidade" de D. A melhor forma de interpretar β_0 e β_1 é considerar os dois casos possíveis, um de cada vez.

Quando $D_i = 0$, o modelo se reduz a $Y_i = \beta_0 + u_i$. Como $\mathbb{E}(u_i \mid D_i) = 0$ pela Hipótese 1, a média condicional de Y_i nesse grupo é

$$\mathbb{E}(Y_i \mid D_i = 0) = \beta_0.$$

Isto é, β_0 é a média populacional dos test scores nos distritos com STR **alta**.

Quando $D_i = 1$, temos $Y_i = \beta_0 + \beta_1 + u_i$, e portanto

$$\mathbb{E}(Y_i \mid D_i = 1) = \beta_0 + \beta_1,$$

a média populacional dos test scores nos distritos com STR **baixa**. A diferença entre as duas é

$$(\beta_0 + \beta_1) - \beta_0 = \beta_1.$$

! O coeficiente da dummy é uma diferença de médias

$$\boxed{\beta_1 = \mathbb{E}(Y_i \mid D_i = 1) - \mathbb{E}(Y_i \mid D_i = 0)}$$

No exemplo do test score:

$$\beta_1 = [\text{média em distritos com } STR < 20] - [\text{média em distritos com } STR \geq 20]\,.$$

E porque β_1 é a diferença entre **médias populacionais**, faz sentido que $\hat{\beta}_1$ seja a diferença entre as **médias amostrais** de Y_i nos dois grupos — o que de fato se demonstra a partir da fórmula do MQO.

Este resultado merece uma pausa, porque ele fecha um círculo aberto no Seção 2. Lá, o efeito causal de um tratamento aleatorizado era estimado pela **diferença entre as médias** dos grupos tratado e de controle. Aqui descobrimos que essa diferença de médias **é** um coeficiente de regressão — o de uma regressão sobre a dummy de tratamento. As duas ferramentas são a mesma ferramenta, e é por isso que a regressão consegue absorver o arcabouço experimental como caso particular. Regressão com dummy e teste de diferença de médias entre dois grupos são o mesmo procedimento escrito de duas maneiras.

5.4.1 Teste de hipótese e IC no caso binário

Estimando o modelo com a dummy nos dados do exemplo, o professor obtém

$$\widehat{TestScore} = 650{,}0 + 7{,}40 \cdot D, \qquad \begin{matrix}(1{,}3) & (1{,}8)\end{matrix}$$

com os erros-padrão entre parênteses, sob a mesma definição de D acima. A leitura é direta:

- o test score médio quando $D = 0$ (isto é, $STR \geq 20$) é 650,0;
- o test score médio quando $D = 1$ (isto é, $STR < 20$) é $650{,}0 + 7{,}4 = 657{,}4$;
- a diferença entre os dois é o coeficiente da dummy, 7,4 pontos.

Em distritos com turmas menores, portanto, o desempenho é 7,4 pontos maior, **em média**.

A diferença entre as médias populacionais dos dois grupos é estatisticamente significante ao nível de 5%? Testamos $H_0 : \beta_1 = 0$ contra $H_1 : \beta_1 \neq 0$:

$$t = \frac{7{,}4 - 0}{1{,}8} = 4{,}04,$$

que excede o valor crítico de 1,96 e está, portanto, na região de rejeição. O intervalo de confiança de 95% para o coeficiente da dummy é

$$7{,}4 \pm 1{,}96 \cdot 1{,}8 \;=\; (3{,}9;\; 10{,}9),$$

que, coerentemente com a dualidade discutida acima, não contém o zero.

Por que $\hat{\beta}_1$ muda ao trocar STR contínuo por dummy

Note que os dois modelos produzem números diferentes — $-2{,}28$ por unidade de STR contra 7,4 pontos entre grupos — e ambos estão corretos, porque respondem a perguntas diferentes. O modelo contínuo estima a inclinação de uma reta ajustada a todo o suporte de STR; a dummy compara duas médias, descartando toda a variação **dentro** de cada grupo. Discretizar um regressor contínuo joga fora informação, e em geral custa precisão; a compensação é não impor linearidade sobre a relação.

5.5 Heterocedasticidade

Chegamos a um assunto que já apareceu de passagem nos dois capítulos anteriores e que agora recebe tratamento próprio. Ele importa precisamente porque **afeta a inferência** — os erros-padrão, os testes t e os intervalos de confiança que acabamos de construir.

Vale recuperar o ponto de partida: a nossa **única** hipótese sobre a distribuição condicional de u_i dado X_i é que sua **média é zero**. Nada dissemos sobre a variância dessa distribuição. Se, além da média zero, a variância condicional **não depende de** X, dizemos que os erros são homocedásticos.

Definição — homocedasticidade e heterocedasticidade

O termo de erro u_i é **homocedástico** se a variância condicional da distribuição de u_i dado X_i é constante para $i = 1, \ldots, n$ e, em particular, **não depende de** X_i:

$$\boxed{\mathrm{Var}(u_i \mid X_i) = \sigma_u^2 \quad \text{(constante)}}$$

Caso contrário, o termo de erro é **heterocedástico**: $\mathrm{Var}(u_i \mid X_i)$ é função de X_i.

Graficamente, a heterocedasticidade aparece como uma nuvem de pontos em forma de funil: a dispersão vertical dos pontos em torno da reta de regressão cresce (ou diminui) conforme se caminha ao longo do eixo X. O exemplo clássico, que o professor projeta, é o da economia do trabalho: salário-hora contra anos de escolaridade na *Current Population Survey*. Entre os trabalhadores com pouca escolaridade, os salários são baixos e pouco dispersos; entre os mais escolarizados, a dispersão é muito maior — há médicos e há biólogos desempregados. A variância do erro claramente cresce com X.

A Figura 7, gerada no laboratório, contrasta os dois casos.

5.5.1 O que a heterocedasticidade não estraga

Aqui é preciso ser preciso, porque a confusão é frequente. Como as hipóteses de MQO **não impõem restrição alguma sobre a variância condicional dos erros**, tudo o que derivamos a partir delas continua valendo:

- os estimadores de MQO continuam **não viesados**: $\mathbb{E}(\hat{\beta}_1) = \beta_1$;
- continuam **consistentes**: $\hat{\beta}_1 \xrightarrow{p} \beta_1$;

- continuam **assintoticamente normais**, com distribuição amostral Normal em amostras grandes.

Isto é: **a heterocedasticidade não viesa o MQO**. O estimador continua acertando o alvo em média. O que ela estraga é outra coisa — a **estimativa da precisão** do estimador.

5.5.2 O que ela estraga: as fórmulas de variância

Sob homocedasticidade, a fórmula da variância de $\hat{\beta}_1$ **simplifica**. Recuperando a forma geral obtida no capítulo anterior e impondo $\text{Var}(u_i \mid X_i) = \sigma_u^2$:

$$\text{Homocedasticidade:} \quad \boxed{\sigma^2_{\hat{\beta}_1} = \frac{s^2_{\hat{u}}}{\sum_{i=1}^{n} (X_i - \bar{X})^2}} \qquad \text{onde} \qquad s^2_{\hat{u}} = \frac{1}{n-2} \sum_{i=1}^{n} \hat{u}_i^2.$$

$$\text{Heterocedasticidade:} \quad \boxed{\sigma^2_{\hat{\beta}_1} = \frac{1}{n} \cdot \frac{\text{Var}[(X_i - \mu_X)\, u_i]}{[\text{Var}(X_i)]^2}}$$

A segunda é a fórmula geral, já apresentada no 1, que **não** supõe nada sobre a variância condicional. A primeira é o caso particular que se obtém dela quando a homocedasticidade vale.

> **A consequência prática**
>
> Se os erros são heterocedásticos e o pesquisador usa a fórmula homocedástica, **toda a parte de inferência da regressão estará incorreta**. Os erros-padrão estarão errados — e, com eles, as estatísticas t, os valores-p e os intervalos de confiança.
> Note a direção do problema: o erro **não** está nas estimativas dos coeficientes, que permanecem não viesadas, mas na avaliação da sua precisão. Reporta-se o número certo com a margem de erro errada, o que pode levar a declarar significante o que não é, ou o contrário.

A solução é usar os **erros-padrão robustos à heterocedasticidade** — os erros-padrão de White (1980), que estimam diretamente a fórmula geral sem supor variância condicional constante. É o que faz a opção `robust` do STATA, presente em todas as saídas do professor, e é o que a função `ep_robusto()` destas notas implementa a partir da fórmula.

> **Por que usar robustos sempre**
>
> A prática recomendada em economia aplicada é usar erros-padrão robustos **por padrão**, e a razão é assimétrica: se os erros são de fato homocedásticos, os robustos são *assintoticamente* equivalentes aos convencionais e nada se perde de relevante em amostras grandes; se são heterocedásticos, os convencionais estão errados e os robustos estão certos. Não há razão para apostar na hipótese mais forte quando o custo de dispensá-la é baixo.
> Em amostras pequenas há uma ressalva: os robustos são consistentes, mas podem ser viesados para baixo em n pequeno, e por isso se usam correções de graus de liberdade — a variante HC1, que é a do STATA e a implementada aqui.

5.5.3 Um exemplo em que quase não faz diferença

O professor apresenta um exemplo instrutivo justamente por ser **anticlimático**. É a regressão de avaliações de curso sobre a beleza do professor (*Beauty in the Classroom*), com $n = 463$ observações. Sem correção:

```
Modelo 1: MQO, usando as observações 1-463
Variável dependente: course_eval
             coeficiente   erro padrão   razão-t   p-valor
   ---------------------------------------------------------
   const       3,99827      0,0253493     157,7    0,0000  ***
   beauty      0,133001     0,0321775       4,133  4,25e-05 ***
```

E com erros-padrão robustos:

```
Modelo 2: MQO, usando as observações 1-463
Variável dependente: course_eval
Erros padrão robustos à heteroscedasticidade
             coeficiente   erro padrão   razão-t   p-valor
   ---------------------------------------------------------
   const      3,99827       0,0253493    157,7     0,0000   ***
   beauty     0,133001      0,0323189      4,115   4,58e-05 ***
```

Os coeficientes são **idênticos** — como devem ser, já que a correção não toca no estimador, apenas no erro-padrão. E os erros-padrão mudam muito pouco: de 0,0322 para 0,0323 no coeficiente de `beauty`, o que move a estatística t de 4,133 para 4,115 e não altera conclusão alguma.

A lição não é que a correção seja dispensável — é que **não se sabe de antemão** se ela fará diferença. Neste conjunto de dados, os erros são aproximadamente homocedásticos e a correção quase não muda nada; em outros, ela muda a conclusão. Como o custo de usá-la é nulo, usa-se sempre.

5.6 O Teorema de Gauss-Markov

O capítulo fecha com o resultado que historicamente serviu de justificativa teórica para o uso do MQO. O professor o apresenta com uma ressalva embutida já no enunciado: o Teorema de Gauss-Markov **era** usado como justificativa. O verbo no passado tem um bom motivo — as hipóteses impostas pelo teorema são tão restritivas que tornam sua validade improvável na prática, o que limita seu alcance como fundamento para o MQO.

> **!** Teorema de Gauss-Markov
>
> Se as **três hipóteses de MQO** são válidas **e os erros são homocedásticos**, então o estimador de MQO $\hat{\beta}_1$ é **BLUE**:
>
> - **B**est — no sentido de **variância mínima**;
> - **L**inear — na classe dos estimadores lineares em Y;
> - **C**onditionally **U**nbiased **E**stimator — condicionalmente não viesado.

Isto é: entre todos os estimadores que são funções lineares de $Y_1, \ldots, Y_n$ e não viesados condicionalmente aos X, o MQO é o de menor variância.

O teorema é um resultado de **eficiência**: não afirma que o MQO é não viesado (isso já vem das Hipóteses 1–3, sem homocedasticidade), mas que, acrescentada a homocedasticidade, ele é o **melhor** dentro de uma classe restrita.

5.6.1 As duas limitações

E é justamente sobre o tamanho dessa classe e a força dessas hipóteses que recaem as duas limitações que o professor enumera.

Primeira: as condições em geral não valem na prática. Em particular, a hipótese de erros homocedásticos **raramente se verifica** em aplicações de economia e finanças. E se os erros são heterocedásticos — como em geral é o caso —, o estimador de MQO **não é mais BLUE**. Existe, sob heterocedasticidade conhecida, um estimador de mínimos quadrados ponderados com variância menor.

Segunda: mesmo quando as condições valem, a classe é restrita. O teorema só compara o MQO com estimadores **lineares** e não viesados. Existem estimadores **não lineares** e não viesados que, sob certas condições, são **mais eficientes** que o MQO.

i Então por que continuamos usando MQO?

A pergunta é legítima, e a resposta não está no Teorema de Gauss-Markov — está no que sobrevive sem ele. Sob as Hipóteses 1–3 **apenas**, sem homocedasticidade e sem qualquer hipótese distribucional, o MQO continua não viesado, consistente e assintoticamente normal. É isso que sustenta a inferência deste capítulo, e é isso que não depende de o mundo ser homocedástico.

O MQO se justifica, portanto, menos por ser ótimo em alguma classe estreita e mais por ser robusto, transparente, computacionalmente trivial e diretamente interpretável como estimador de uma média condicional. As alternativas que o dominam em eficiência exigem, em troca, conhecer a estrutura da heterocedasticidade — conhecimento que raramente se tem.

O que **não** se resolve por essa via é o problema central do curso: nenhuma propriedade de eficiência salva o MQO se $\mathbb{E}(u_i \mid X_i) \neq 0$. Um estimador viesado com variância mínima continua sendo um estimador viesado, e é contra o viés — não contra a ineficiência — que o resto do curso vai trabalhar.

5.7 Laboratório computacional em R

O laboratório deste capítulo tem cinco partes. Reproduz os testes t e os intervalos de confiança da regressão do exemplo; verifica por simulação o que significa "95% de confiança"; ilustra graficamente homocedasticidade contra heterocedasticidade; mostra numericamente o que acontece com a inferência quando se usa a fórmula errada; e reproduz a equivalência entre regressão com dummy e diferença de médias.

Retomamos os mesmos dados simulados do capítulo anterior — com a mesma advertência: eles **não** são os dados originais de Stock e Watson (2020), e sim uma calibração que reproduz a ordem de grandeza dos resultados do slide.

```
set.seed(5490)
n_d     <- 420
beta0   <- 698.9
beta1   <- -2.28
sigma_u <- 18.55

STR       <- rnorm(n_d, mean = 19.64, sd = 1.89)
u         <- rnorm(n_d, mean = 0,     sd = sigma_u)
TestScore <- beta0 + beta1 * STR + u
ca <- data.frame(STR = STR, TestScore = TestScore)

reg <- lm(TestScore ~ STR, data = ca)
round(coef(reg), 4)
```

```
(Intercept)         STR
   698.8918     -2.2682
```

5.7.1 (a) A tabela completa de inferência

A função `tabela_reg()`, definida no preâmbulo destas notas, monta a tabela que o STATA imprime: coeficiente, erro-padrão robusto, estatística t, valor-p e os limites do intervalo de confiança de 95%.

```
round(tabela_reg(reg, robusto = TRUE), 4)
```

```
                coef     ep       t p   ic_inf   ic_sup
(Intercept) 698.8918 8.9701 77.9138 0 681.2597 716.5238
STR          -2.2682 0.4587 -4.9446 0  -3.1699  -1.3665
```

Cada coluna é um objeto deste capítulo. Para o coeficiente de `STR`: a estatística t testa $H_0 : \beta_1 = 0$ e é a razão entre a estimativa e o erro-padrão; o valor-p é $2\,F_t(-|t|)$; e o intervalo de confiança é $\hat{\beta}_1 \pm t_{0{,}975} \cdot \mathrm{SE}(\hat{\beta}_1)$. Compare com a saída do STATA reproduzida acima — a correspondência é linha a linha, com as diferenças esperadas por se tratar de dados simulados.

Vale conferir "na mão" que as três colunas se produzem umas às outras:

```
b1  <- coef(reg)[2]
se1 <- ep_robusto(reg)[2]
t1  <- b1 / se1
gl  <- df.residual(reg)

c(beta1_hat = b1,
  ep        = se1,
```

```
  t         = t1,
  valor_p   = 2 * pt(-abs(t1), gl),
  ic_inf    = b1 - qt(0.975, gl) * se1,
  ic_sup    = b1 + qt(0.975, gl) * se1) |> round(4)
```

```
beta1_hat.STR        ep.STR         t.STR   valor_p.STR    ic_inf.STR
      -2.2682        0.4587       -4.9446        0.0000       -3.1699
   ic_sup.STR
      -1.3665
```

Note que uso o valor crítico da distribuição t com $n-2$ graus de liberdade (`qt(0.975, 418)`), e não o 1,96 da Normal. Com 418 graus de liberdade a diferença é irrelevante — o valor crítico é praticamente 1,966 —, mas é o que o STATA faz e o que mantém a tabela coerente com a coluna `P>|t|`.

```
c(normal = qnorm(0.975), t_418 = qt(0.975, gl), t_20 = qt(0.975, 20)) |> round(4)
```

```
normal  t_418   t_20
1.9600 1.9657 2.0860
```

Com n pequeno a diferença deixa de ser irrelevante: com 20 graus de liberdade o valor crítico sobe para 2,086, e ignorá-lo produziria intervalos 6% mais estreitos do que deveriam ser.

5.7.2 (b) O efeito predito de $\Delta STR = -2$

```
delta   <- -2
ic_b1   <- c(b1 - qt(0.975, gl) * se1, b1 + qt(0.975, gl) * se1)
ic_efeito <- sort(ic_b1 * delta)

c(efeito_pontual = b1 * delta,
  ic_inf = ic_efeito[1],
  ic_sup = ic_efeito[2]) |> round(4)
```

```
efeito_pontual.STR         ic_inf.STR         ic_sup.STR
            4.5365             2.7331             6.3399
```

O `sort()` não é cosmético: como $\Delta x = -2$ é negativo, a multiplicação **inverte** a ordem dos extremos, exatamente como discutido no texto. Sem reordenar, o "limite inferior" sairia maior que o superior.

5.7.3 (c) O que significa "95% de confiança"

Esta é a parte central do laboratório. Vamos construir o experimento de amostras repetidas que a interpretação formal do intervalo descreve: geramos 5.000 amostras do mesmo processo, calculamos o IC de 95% em cada uma e contamos em quantas delas o intervalo contém o **verdadeiro** $\beta_1 = -2{,}28$.

```
set.seed(2026)
n_rep <- 5000
n_am  <- 420

cobre <- logical(n_rep)
largura <- numeric(n_rep)

for (r in seq_len(n_rep)) {
  Xr <- rnorm(n_am, mean = 19.64, sd = 1.89)
  Yr <- beta0 + beta1 * Xr + rnorm(n_am, mean = 0, sd = sigma_u)
  mr <- lm(Yr ~ Xr)

  br  <- coef(mr)[2]
  sr  <- ep_robusto(mr)[2]
  tc  <- qt(0.975, df.residual(mr))

  li <- br - tc * sr
  ls <- br + tc * sr
  cobre[r]   <- (li < beta1) && (beta1 < ls)
  largura[r] <- ls - li
}

c(cobertura_empirica = mean(cobre),
  nominal            = 0.95,
  largura_media      = mean(largura)) |> round(4)
```

```
cobertura_empirica            nominal      largura_media
            0.9478             0.9500             1.8770
```

A cobertura empírica fica muito próxima dos 95% nominais. **É isso, literalmente, que a frase "intervalo de 95% de confiança" quer dizer**: não que este intervalo específico tenha 95% de chance de conter β_1, mas que o *procedimento* que o gerou acerta em 95% das amostras.

A Figura 6 mostra os 100 primeiros intervalos dessa simulação.

```
set.seed(2026)
n_mostrar <- 100
ics <- data.frame(rep = seq_len(n_mostrar), li = NA_real_, ls = NA_real_, b = NA_real_)

for (r in seq_len(n_mostrar)) {
  Xr <- rnorm(n_am, mean = 19.64, sd = 1.89)
```

```
  Yr <- beta0 + beta1 * Xr + rnorm(n_am, mean = 0, sd = sigma_u)
  mr <- lm(Yr ~ Xr)
  br <- coef(mr)[2]; sr <- ep_robusto(mr)[2]; tc <- qt(0.975, df.residual(mr))
  ics$li[r] <- br - tc * sr; ics$ls[r] <- br + tc * sr; ics$b[r] <- br
}
ics$acerta <- ics$li < beta1 & beta1 < ics$ls

ggplot(ics, aes(y = rep, x = b, xmin = li, xmax = ls, colour = acerta)) +
  geom_vline(xintercept = beta1, colour = cz, linewidth = 0.6) +
  geom_errorbar(orientation = "y", width = 0, linewidth = 0.4, alpha = 0.9) +
  geom_point(size = 0.7) +
  scale_colour_manual(values = c(`TRUE` = az, `FALSE` = vm),
                      labels = c(`TRUE` = "contém", `FALSE` = "não contém")) +
  labs(x = expression(hat(beta)[1]), y = "réplica")
```

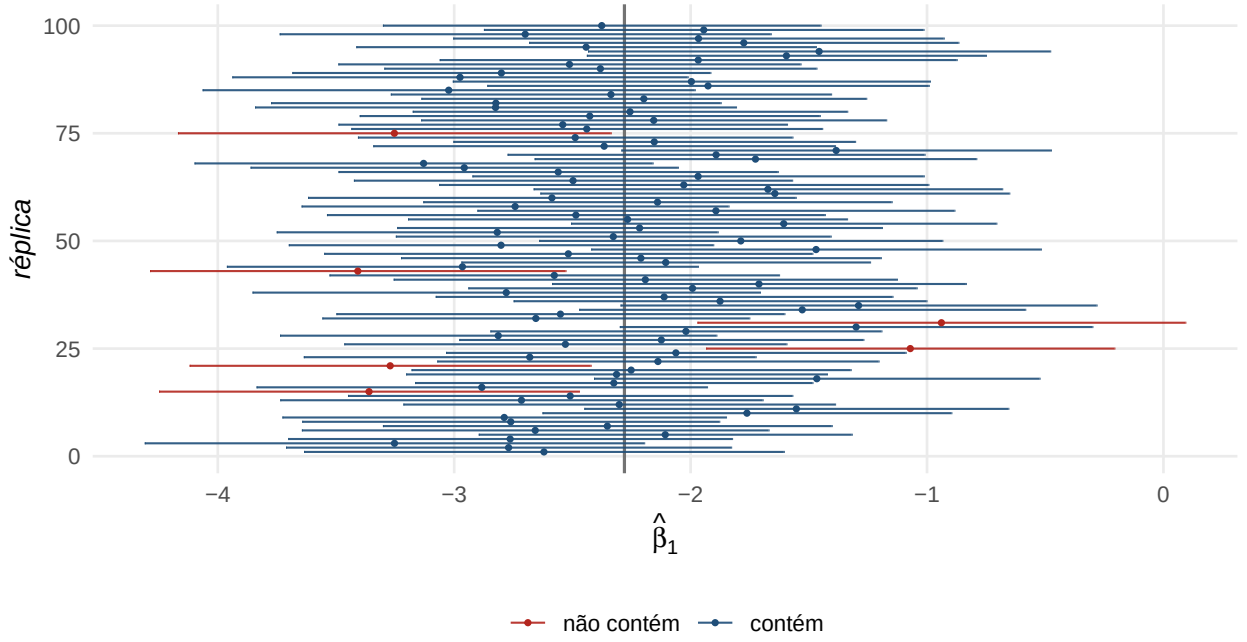

Figura 6: Cem intervalos de confiança de 95% para β_1, cada um construído a partir de uma amostra diferente do mesmo processo gerador. A linha vertical marca o verdadeiro $\beta_1 = -2{,}28$. Os intervalos em azul contêm o parâmetro; os em vermelho, não. O que varia de amostra para amostra é o intervalo, não o parâmetro — e é sobre a frequência de acerto do procedimento, ao longo de amostras repetidas, que a confiança de 95% faz uma afirmação.

Alguns dos 100 intervalos exibidos erram o alvo, em proporção compatível com os 5% esperados — em uma amostra de apenas 100 réplicas, o próprio número de erros é aleatório. Nenhum dos que erram é "o intervalo errado" em algum sentido especial: todos foram construídos pela mesma regra, a partir de amostras igualmente legítimas. Na prática temos **uma** amostra e nunca sabemos se caímos entre os azuis ou entre os vermelhos.

5.7.4 (d) Homocedasticidade contra heterocedasticidade

```
set.seed(11)
n_h <- 300
Xh  <- runif(n_h, 1, 10)

d_homo   <- data.frame(X = Xh, Y = 2 + 1.5 * Xh + rnorm(n_h, 0, 2),
                       tipo = "Homocedástico: V(u|X) constante")
d_hetero <- data.frame(X = Xh, Y = 2 + 1.5 * Xh + rnorm(n_h, 0, 0.45 * Xh),
                       tipo = "Heterocedástico: V(u|X) cresce com X")

rbind(d_homo, d_hetero) |>
  ggplot(aes(X, Y)) +
  geom_point(alpha = 0.45, size = 0.9, colour = az) +
  geom_abline(intercept = 2, slope = 1.5, colour = vm, linewidth = 0.7) +
  facet_wrap(~tipo)
```

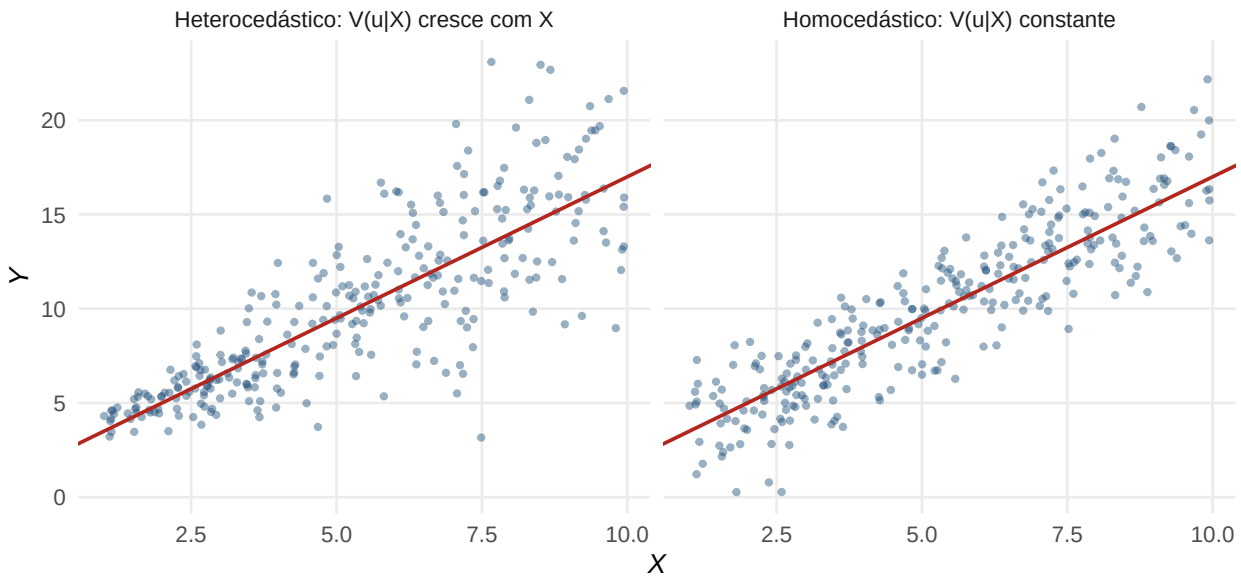

Figura 7: Erros homocedásticos (esquerda) e heterocedásticos (direita), com a mesma reta de regressão populacional. À esquerda, a dispersão vertical dos pontos em torno da reta é a mesma em todo o suporte de X. À direita, ela cresce com X — a nuvem tem forma de funil. Em ambos os casos o MQO estima a reta corretamente; o que muda é a precisão dessa estimativa em cada região do suporte, e portanto a fórmula correta do erro-padrão.

5.7.5 (e) A inferência quando se usa a fórmula errada

Aqui está a consequência prática, medida. Geramos dados **deliberadamente heterocedásticos**, com $\mathrm{Var}(u_i \mid X_i) = (0{,}5\,X_i)^2$, e comparamos três coisas ao longo de 3.000 réplicas: o desvio-padrão

empírico de $\hat{\beta}_1$ (que é a verdade contra a qual se julga), a média dos erros-padrão **homocedásticos** e a média dos **robustos**.

```
set.seed(99)
n_rep2 <- 3000
n_am2  <- 200

b_vec  <- numeric(n_rep2)
ep_hom <- numeric(n_rep2)
ep_rob <- numeric(n_rep2)
cob_hom <- logical(n_rep2)
cob_rob <- logical(n_rep2)

b1_true <- 1.5

for (r in seq_len(n_rep2)) {
  Xr <- runif(n_am2, 1, 10)
  Yr <- 2 + b1_true * Xr + rnorm(n_am2, 0, 0.5 * Xr)   # heterocedastico
  mr <- lm(Yr ~ Xr)
  tc <- qt(0.975, df.residual(mr))

  b_vec[r]  <- coef(mr)[2]
  ep_hom[r] <- summary(mr)$coefficients[2, 2]
  ep_rob[r] <- ep_robusto(mr)[2]

  cob_hom[r] <- abs(b_vec[r] - b1_true) < tc * ep_hom[r]
  cob_rob[r] <- abs(b_vec[r] - b1_true) < tc * ep_rob[r]
}

round(c(dp_empirico_de_beta1 = sd(b_vec),
        media_ep_homocedastico = mean(ep_hom),
        media_ep_robusto       = mean(ep_rob)), 4)
```

```
  dp_empirico_de_beta1 media_ep_homocedastico       media_ep_robusto
                0.0893                 0.0828                 0.0881
```

O desvio-padrão empírico de $\hat{\beta}_1$ é a referência: é a variabilidade que o estimador de fato tem. O erro-padrão robusto o reproduz de perto; o homocedástico **subestima** a variabilidade verdadeira. E a consequência aparece na cobertura dos intervalos:

```
round(c(cobertura_com_ep_homocedastico = mean(cob_hom),
        cobertura_com_ep_robusto       = mean(cob_rob),
        nominal                        = 0.95), 4)
```

```
cobertura_com_ep_homocedastico       cobertura_com_ep_robusto
                        0.9323                         0.9433
```

```
nominal
 0.9500
```

O intervalo construído com o erro-padrão homocedástico, que se anuncia como de 95%, cobre o parâmetro em menos que 95% das amostras — é **estreito demais**. Quem o utiliza rejeita H_0 com mais frequência do que o nível declarado, isto é, comete erros do Tipo I acima do α que julga ter fixado. O intervalo robusto fica sensivelmente mais próximo do nível nominal.

A cobertura robusta também não bate exatamente os 95%, e vale entender por quê: com $n = 200$, o erro-padrão robusto é ele próprio **estimado** com ruído, e a correção HC1 não elimina inteiramente o viés de amostra finita. A garantia dos erros-padrão robustos é **assintótica**; o que a simulação mostra é a direção e a ordem de grandeza da correção, não uma promessa exata em n pequeno. O exercício 7(d) pede refazer isto com $n = 30$, onde a discrepância é bem maior.

Note também que a média de $\hat{\beta}_1$ ao longo das réplicas continua no valor verdadeiro, confirmando que a heterocedasticidade não viesa o estimador:

```r
c(media_beta1_hat = mean(b_vec), beta1_verdadeiro = b1_true) |> round(4)
```

```
 media_beta1_hat beta1_verdadeiro
          1.4961           1.5000
```

5.7.6 (f) Regressão com dummy é diferença de médias

Por fim, a verificação numérica da equivalência estabelecida no texto. Criamos a dummy $D = 1\{STR < 20\}$ nos dados do exemplo e comparamos o coeficiente da regressão com a diferença entre as médias amostrais dos dois grupos.

```r
ca$D <- as.numeric(ca$STR < 20)

reg_d <- lm(TestScore ~ D, data = ca)

media_D1 <- mean(ca$TestScore[ca$D == 1])   # STR < 20
media_D0 <- mean(ca$TestScore[ca$D == 0])   # STR >= 20

round(c(intercepto_regressao = coef(reg_d)[1],
        media_grupo_D0       = media_D0,
        coef_dummy           = coef(reg_d)[2],
        diferenca_de_medias  = media_D1 - media_D0), 4)
```

```
intercepto_regressao.(Intercept)                 media_grupo_D0
                        649.8960                       649.8960
                     coef_dummy.D              diferenca_de_medias
                          7.8604                         7.8604
```

O intercepto **é** a média do grupo $D = 0$ e o coeficiente da dummy **é** a diferença entre as médias — não aproximadamente, mas até o último dígito. A regressão sobre uma dummy é uma comparação de médias escrita em outra notação.

E a inferência sobre esse coeficiente é a mesma de sempre:

```
round(tabela_reg(reg_d, robusto = TRUE), 4)
```

```
                coef     ep        t p   ic_inf   ic_sup
(Intercept) 649.8960 1.4597 445.2113 0 647.0266 652.7654
D             7.8604 1.8651   4.2145 0   4.1943  11.5264
```

Compare com o teste t de diferença de médias entre os dois grupos, com variâncias desiguais (Welch), que é o procedimento que um curso de estatística aplicaria ao mesmo problema:

```
tt <- t.test(TestScore ~ D, data = ca)
round(c(t_welch = unname(tt$statistic),
        p_welch = tt$p.value,
        t_regressao_robusto = tabela_reg(reg_d)[2, "t"],
        p_regressao_robusto = tabela_reg(reg_d)[2, "p"]), 4)
```

```
           t_welch            p_welch t_regressao_robusto p_regressao_robusto
           -4.2141             0.0000              4.2145               0.0000
```

As estatísticas são praticamente idênticas (o sinal difere apenas pela ordem em que cada função toma a diferença entre os grupos). A pequena discrepância vem de a correção de graus de liberdade de Welch não ser exatamente a do HC1 — mas a conclusão é a mesma, como tem de ser: é o mesmo teste.

5.8 Exercícios

1. **Os dois erros.** Um órgão regulador testa se um medicamento tem efeito, com H_0: "o medicamento não tem efeito". (a) Descreva em palavras o erro do Tipo I e o do Tipo II neste contexto, e diga qual deles lhe parece mais custoso socialmente. (b) O regulador decide reduzir α de 5% para 1%, mantendo o tamanho da amostra. O que acontece com β e com o poder? (c) Um laboratório concorrente, que quer demonstrar que o medicamento **não** funciona, tem incentivo a escolher qual tamanho de amostra? Relacione com a advertência do professor sobre poder. (d) Reescreva a nula de modo a inverter os papéis dos dois erros, e discuta quando essa formulação seria preferível.

2. **Teste sobre a média.** Refaça o exemplo do texto ($\sigma = 0{,}1$, $n = 16$, $\bar{x} = 5{,}038$) supondo agora a alternativa **bicaudal** $H_1 : \mu \neq 5$. (a) Qual é o novo valor crítico a 5%? (b) Qual é o valor-p? (c) A conclusão muda? (d) Explique por que o valor-p bicaudal é exatamente o dobro do unicaudal neste caso, e sob que condição essa relação vale.

3. **Lendo uma tabela de regressão.** Uma regressão de salário sobre anos de escolaridade, com $n = 1.200$, produziu $\hat{\beta}_1 = 0{,}087$ com erro-padrão robusto 0,011. (a) Calcule a estatística t para $H_0 : \beta_1 = 0$ e conclua a 1%. (b) Construa o IC de 95% e o de 99%; qual é mais largo, e por quê? (c) Teste $H_0 : \beta_1 = 0{,}10$ contra a alternativa bicaudal — o resultado é coerente com o IC de 95% que você construiu?

   (d) Um pesquisador afirma que "o retorno da escolaridade é de 8,7% e o valor-p é praticamente zero, logo o efeito é grande". Avalie as duas afirmações separadamente.

4. **A dualidade.** Mostre algebricamente que o teste bicaudal de $H_0 : \beta_1 = \beta_{1,0}$ ao nível α não rejeita **se e somente se** $\beta_{1,0}$ pertence ao intervalo de confiança de nível $1 - \alpha$. (b) Use essa equivalência para responder, sem calcular nenhuma estatística t: dado o IC de 95% $(-3{,}30;\ -1{,}26)$ do exemplo, quais dos valores $\{-4;\ -3;\ -1{,}5;\ 0;\ 1\}$ seriam rejeitados a 5%? (c) A dualidade vale para testes unicaudais? Justifique.

5. **Efeitos preditos.** Com $\hat{\beta}_1 = -2{,}28$ e $\mathrm{SE}(\hat{\beta}_1) = 0{,}52$: (a) construa o IC de 95% para o efeito de $\Delta STR = -5$. (b) Explique por que os extremos se invertem quando $\Delta x < 0$. (c) Suponha agora que se queira o IC para o efeito de $\Delta STR = -2$ **em pontos percentuais** do desvio-padrão dos test scores (≈ 19 pontos). Obtenha-o. (d) Por que não é necessário conhecer $\mathrm{SE}(\hat{\beta}_0)$ para nada disso?

6. **Regressor binário.** Considere $Y_i = \beta_0 + \beta_1 D_i + u_i$ com D_i binária.

   (a) Partindo da fórmula do MQO $\hat{\beta}_1 = \frac{\sum (D_i - \bar{D})(Y_i - \bar{Y})}{\sum (D_i - \bar{D})^2}$, demonstre que $\hat{\beta}_1 = \bar{Y}_1 - \bar{Y}_0$, as médias amostrais nos dois grupos.
   (b) Mostre que $\hat{\beta}_0 = \bar{Y}_0$. (c) Suponha que se defina a dummy ao contrário, $\tilde{D} = 1 - D$: o que acontece com $\hat{\beta}_0$, $\hat{\beta}_1$, a estatística t e o R^2? (d) O que aconteceria se incluíssemos **ambas**, D e $\tilde{D}$, além do intercepto?

7. **Heterocedasticidade.** (a) Explique por que a heterocedasticidade **não** viesa $\hat{\beta}_1$, apontando em que passo da demonstração de não-viés a variância condicional dos erros seria usada — e verificando que ela não é. (b) Na simulação da seção (e) do laboratório, o erro-padrão homocedástico subestimou a variabilidade verdadeira. Isso é sempre assim, ou existe padrão de heterocedasticidade que produza o contrário? Investigue por simulação com $\mathrm{Var}(u_i \mid X_i)$ **decrescente** em X_i. (c) Por que os erros-padrão robustos são preferidos "por padrão", mesmo quando os erros podem ser homocedásticos? (d) O que muda se $n = 30$ em vez de $n = 200$?

8. **Gauss-Markov.** (a) Enuncie precisamente as hipóteses do teorema e identifique qual delas **não** está entre as três hipóteses de MQO. (b) Explique a diferença entre "não viesado" e "BLUE", e por que a heterocedasticidade destrói o segundo mas não o primeiro. (c) O teorema afirma que o MQO tem variância mínima entre estimadores lineares e não viesados. Dê um exemplo de estimador **linear e viesado** com variância menor que a do MQO — isso contradiz o teorema? (d) À luz das duas limitações discutidas no texto, como você justificaria o uso do MQO a um cético?

9. **Simulação de poder.** Adapte o código da seção (c) do laboratório para estimar o **poder** do teste de $H_0 : \beta_1 = 0$. (a) Fixe $\beta_1 = -0{,}5$, $n = 100$, e estime, ao longo de 2.000 réplicas, a fração de amostras em que H_0 é rejeitada a 5% — esse é o poder contra essa alternativa. (b) Repita para $n \in \{50, 100, 400, 1600\}$ e descreva como o poder evolui. (c) Repita para $\beta_1 \in \{-0{,}1,\ -0{,}5,\ -1,\ -2\}$ com $n = 100$ fixo, e trace a **curva de poder**. (d) Verifique que,

quando $\beta_1 = 0$, a fração de rejeições é aproximadamente α — e explique por que isso tem de ser assim.

Referências

Angrist, Joshua D., e Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Angrist, Joshua D., e Jörn-Steffen Pischke. 2015. *Mastering 'Metrics: The Path from Cause to Effect*. Princeton University Press.

Bombardier, Claire, Loren Laine, Alise Reicin, et al. 2000. "Comparison of Upper Gastrointestinal Toxicity of Rofecoxib and Naproxen in Patients with Rheumatoid Arthritis". *New England Journal of Medicine* 343 (21): 1520–28.

Borba, Mayla G. S., Fernando F. A. Val, Vanderson S. Sampaio, et al. 2020. "Effect of High vs Low Doses of Chloroquine Diphosphate as Adjunctive Therapy for Patients Hospitalized with Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Infection". *JAMA Network Open* 3 (4): e208857.

Doll, Richard, e A. Bradford Hill. 1954. "The Mortality of Doctors in Relation to Their Smoking Habits". *British Medical Journal* 1 (4877): 1451–55.

Stock, James H., e Mark W. Watson. 2020. *Introduction to Econometrics*. 4º ed. Pearson.

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity". *Econometrica* 48 (4): 817–38.

Wooldridge, Jeffrey M. 2016. *Introductory Econometrics: A Modern Approach*. 7º ed. Cengage Learning.

Wynder, Ernest L., e Evarts A. Graham. 1950. "Tobacco Smoking as a Possible Etiologic Factor in Bronchiogenic Carcinoma". *Journal of the American Medical Association* 143 (4): 329–36.